

UNCLASSIFIED

AD 297 248

*Reproduced
by the*

**ARMED SERVICES TECHNICAL INFORMATION AGENCY
ARLINGTON HALL STATION
ARLINGTON 12, VIRGINIA**



UNCLASSIFIED

NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.

CATALOGED BY ASTIA
AS AD NO. 297248

RADC-TDR-62-598

63-2-5
297248

COPY NO. 41

30 November 1962

Final Report

OPTIMUM SPEECH SIGNAL MAPPING TECHNIQUES

W. B. Floyd

LITTON SYSTEMS, INC.

Data Systems Division

Communication Sciences Laboratory

221 Crescent Street

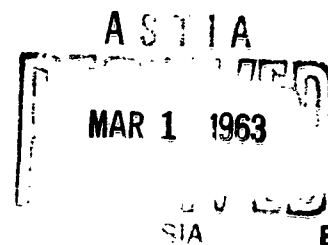
Waltham 54, Massachusetts

Contract AF30(602)-2641

Program Element Code No. 62405484

Prepared for
Rome Air Development Center
Air Force Systems Command
United States Air Force

Griffiss Air Force Base
New York



Qualified requestors may obtain copies of this report from the ASTIA Document Service Center, Dayton 2, Ohio. ASTIA Services for the Department of Defense contractors are available through the "Field of Interest Register" on a "need-to-know" certified by the cognizant military agency of their project or contract.

Final Report
Optimum Speech Signal Mapping Techniques
W. B. Floyd

Litton Systems, Inc.
Communication Sciences Laboratory
Data Systems Division
Waltham 54, Massachusetts

Contract Number AF30(602)-2641
Program Element Code No. 62405484
760D - Project 4027, Task 402704
Electronic Systems Division (RADC)
E. Chapin, RAWER

Prepared for
Rome Air Development Center
Air Force Systems Command
United States Air Force
Griffiss Air Force Base
New York

Qualified requestors may obtain copies of this report from the ASTIA Document Service Center, Dayton 2, Ohio. ASTIA Services for the Department of Defense contractors are available through the "Field of Interest Register" on a "need-to-know" certified by the cognizant military agency of their project or contract.

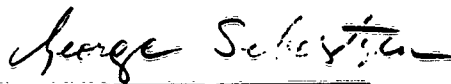
Optimum Speech Signal Mapping Techniques

Contract AF30(602)-2641

30 November 1962

Prepared by
William Floyd

Approved by



George Sebestyen
Technical Director



John Gerdes
Assistant Manager

LITTON SYSTEMS, INC.
DATA SYSTEMS DIVISION

Foreword

This report has been prepared by the Communication Sciences Laboratory within the Data Systems Division of Litton Systems, Inc., a division of Litton Industries. The work reported here has been performed over a period of 9 months, under Contract Number AF30(602)-2641, as Task Number 402704 of Project Number 4027, entitled "Optimum Speech Signal Mapping Techniques". This project has been completed under the direction of the I and EW Directorate within the Rome Air Development Center.

Several individuals within the Communication Sciences Laboratory have made major contributions to the development of speech processing techniques reported here. Experimental speech processing equipment has been designed and constructed under the guidance of Mr. Arthur Crooke; computer programs have been prepared by Mr. Paul Connolly and Miss Helen O'Shea; and the transcription techniques were developed by Mr. William Floyd, Miss Alice Hartley, and Dr. George Sebestyen. Mr. Floyd has served as Project Engineer and Dr. Sebestyen has provided technical guidance for the entire program.

In addition to the other individuals in the Communication Sciences Laboratory who contributed to this work, thanks are due to Mr. Caldwell Smith of the Air Force Cambridge Research Laboratories for his generous contribution of quantized speech data and synthesized speech recordings during the early phase of the study.

ABSTRACT

The problem of representing speech signals in a format which will facilitate automatic speech transcription has been investigated. The method of representation selected for experimental study involves the transformation of speech signals into sequences of periodically sampled outputs of speech parameter extractors, i. e., devices designed to extract clues from speech signals which will serve to identify the language element being uttered. Automatic extractors have been constructed and data has been collected to ascertain the degree to which speech sounds can be identified properly, using several parameters reflecting the location of formants and spectral shape information.

Methods of completing the transformation, or transcription, of speech into sequences of language elements suitable for presentation to a human reader have also been investigated. Test results indicate that the most easily instrumented transcription methods can be expected to yield readable transcriptions from the use of a small number of speech parameters.

TABLE OF CONTENTS

1.	INTRODUCTION	1
2.	THEORETICAL BASIS FOR AUTOMATIC SPEECH TRANSCRIPTION	5
	2.1 Selection of Language Elements	5
	2.2 Selection of Speech Parameters	9
	2.3 Pattern Recognition Methods	19
3.	SPEECH REPRESENTATION IN PARAMETER SPACE	26
	3.1 Speech Parameter Extraction	26
	3.1.1 Description of Equipment	28
	3.1.2 Parameter Quantization	40
	3.2 Distribution of Speech Sounds in Parameter Space	41
	3.2.1 Parameter Space Usage	45
	3.2.2 Overlap Between Speech Sounds in Parameter Space	52
4.	SPEECH TRANSCRIPTION AND WORD RECOGNITION TECHNIQUES	62
	4.1 Speech Transcription Methods	62
	4.2 Word Recognition Methods	70
5.	CONCLUSIONS AND RECOMMENDATIONS	75
	APPENDIX I	79
	APPENDIX II	82
	APPENDIX III	90
	Reference List	95

List of Figures

<u>Number</u>		<u>Page</u>
1	Two Basic Transformations in the Representation of Speech	3
2	Hypothetical Distribution of Two Speech Sounds in a Two-Dimensional Parameter Space	15
3	Hypothetical Multi-Modal Distribution of Two Speech Sounds in a Two-Dimensional Parameter Space	22
4	Block Diagram of Approach to Parameter Extraction and Speech Signal Mapping	29
5	Block Diagram of Experimental Speech Processing Equipment	30
6	Litton Vocoder Spectrum Analyzer Filter Bank Response	31
7	Spectrum Peak Picker	33
8	Formant Indications Provided by a Spectrogram and an Automatic Peak Picker for the Word "ONE"	34
9	Outputs of First and Second Spectrum Moment Extractors for Sinewave Inputs at Different Frequencies	36
10	Spectrum Peak Picker Output Representation of the Spoken Word "ASK"	39
11	Reference Library Structure by Number of Spectrum Peaks	51
12	Spectral Peak Profiles of Vowel Sounds for Speaker Number One	59
13	Spectral Peak Profiles of Vowel Sounds for Speaker Number Two	60
14	Spectral Peak Profiles of Vowel Sounds for Speaker Number Three	61
15	Recognition of the Words "ART" and "TAR"	74
16	Flow Chart for Spectrum Peak Picking Program	82
17	Three-Bit Quantized and Peak-Picked Representation of the Spoken Word "ONE"	83
18	Flow Chart for SMREF	94
19	Flow Chart for Matching (subroutine for SMREF)	95
20	Output Section for SMREF	96

List of Tables

<u>Number</u>		<u>Page</u>
1	Phonetic Alphabet	10
2	Correspondence Between Speech Features and Phonemes	13
3	List of Speech Parameters Investigated	27
4	Segmentation of "Two Three" with the Parameter ΔS	37
5	Speech Parameter Quantization	40
6	Three Word Lists Employed for Vowel Sounds	42
7	Representation of the Word "NECK" in Parameter Space	43
8	Binary Representations of Speech Parameters	44
9	Number of Samples and Patterns Obtained from Three Speakers and Vowel Sounds	46
10	Complete Histogram for the Sound EE(i) in Peak Space (Speaker Number One).	48
11	Fraction of Vowel Samples Covered by the Ten Most Frequently Occurring Patterns	50
12	Estimated Relative Frequency of Correct and Misclassification of Vowel Sounds for Speaker Number One, and Two Parameter Spaces	54
13	Estimated Relative Frequency of Correct and Misclassification of Vowel Sounds for Speaker Number Two, and Two Parameter Spaces	55
14	Estimated Relative Frequency of Correct and Misclassification of Vowel Sounds for Speaker Number Three, and Two Parameter Spaces	56
15	Number of Different Patterns with p Peaks which Ever Arise from k Speech Sounds	57
16	Exact Match Transcription of Test Word List for Speaker Number One	67
17	Exact Match Transcription of Test Word List for Speaker Number Two	68
18	Exact Match Transcription of Test Word List for Speaker Number Three	69
19	Test Word List	73
20	Relative Frequency of Occurrence of Sounds	76

I INTRODUCTION

The general problem with which this study has been concerned is that of determining efficient methods of transforming speech signals into sequences of language elements suitable for presentation either to a human or to a machine. In the case of a human recipient, the transformed speech should convey the same information as would be possible through the use of a human stenographer and typist. Thus, a machine designed to implement the speech transformation methods might be called a "phonetic typewriter".* If the set of language elements into which the speech signals are transformed consists of a phonetic alphabet, then such a machine could be designated more accurately as an automatic speech transcriber.

An automatic speech transcription capability is applicable to essentially any communications problem involving (a) human speech as an information source or relay, (b) a temporary or permanent storage requirement, and (c) a need for rapid human assimilation of the information. A person can read printed matter at the rate of many hundreds of words per minute; however, a speaker generates information at a considerably lower rate. To achieve the higher rate of assimilation, speech must be converted by some means to printed English. All current methods of conversion from speech to printed text involve either at least one additional person or a considerable delay or both. An automatic speech transcription device would replace the extra individual as well as eliminate or reduce significantly the transcription delay.

In addition to and perhaps more important than its utility as a transcriber of text, a speech transcription technique inherently carries with it the capability for voice control of machines. Thus, for instance, instead of the depression of keys, pedals, buttons and the like as a means of feeding information into a computer, a speech transcriber with a word recognition unit could be used to program as well as insert data into the computer. Thus, in many applications involving the transfer of human-generated instructions to machines, a speech transcriber can serve to relieve the human from the burden of having to learn new, and usually relatively slow methods of communication, by allowing the use of a natural method: a spoken language.

*See [6]. Each reference in this report is indicated by a number enclosed in brackets. The Reference List at the end of the report identifies the numbers with full descriptions of the references.

The specific purpose of the current project under Contract No. AF30(602)-2641 is to find an optimum format for representation of speech signals to facilitate automatic speech transcription. It is desired that the derived representation be optimized with respect to accuracy of representation, storage requirements, and ease of implementation.

The derivation of such a representation requires that suitable measurable speech signal properties, or parameters, be found which serve to preserve the linguistic information in speech, and also serve as suitable inputs to a language element recognizer. The extraction of these parameters may be regarded as a transformation, or mapping, from "speech signal space" to "parameter space". As depicted in Figure 1, this transformation, T_1 , is to be followed by another T_2 , which would complete the conversion of speech to readable form by mapping the elements of parameter space into a space of language elements. Although the primary purpose of this study has been to investigate the initial transformation, T_1 , results have also been obtained for a few methods of completing the transcription of speech, i. e., performing T_2 using a phonetic alphabet as the language element space.

The approach taken on this project has been to investigate first the accuracy of representation of speech attainable with the simplest form of implementation and minimum extracted speech data. Through the systematic augmentation of extracted speech parameters and refinement of recognition methods, the following results have been assured:

- 1) Speech representation accuracy will always improve as further effort is expended, and
- 2) Reliable relationships between representation accuracy, information storage requirements, and equipment simplicity will be obtained, from which a judgment as to an optimum combination can be made.

In Section 2 of this report the theory underlying these transcription methods is reviewed. The basis for using a phonetic alphabet, rather than some other set of language elements, is presented along with discussions of the problem of selecting appropriate speech parameters and methods of processing parameter values to recognize letters in the phonetic alphabet.

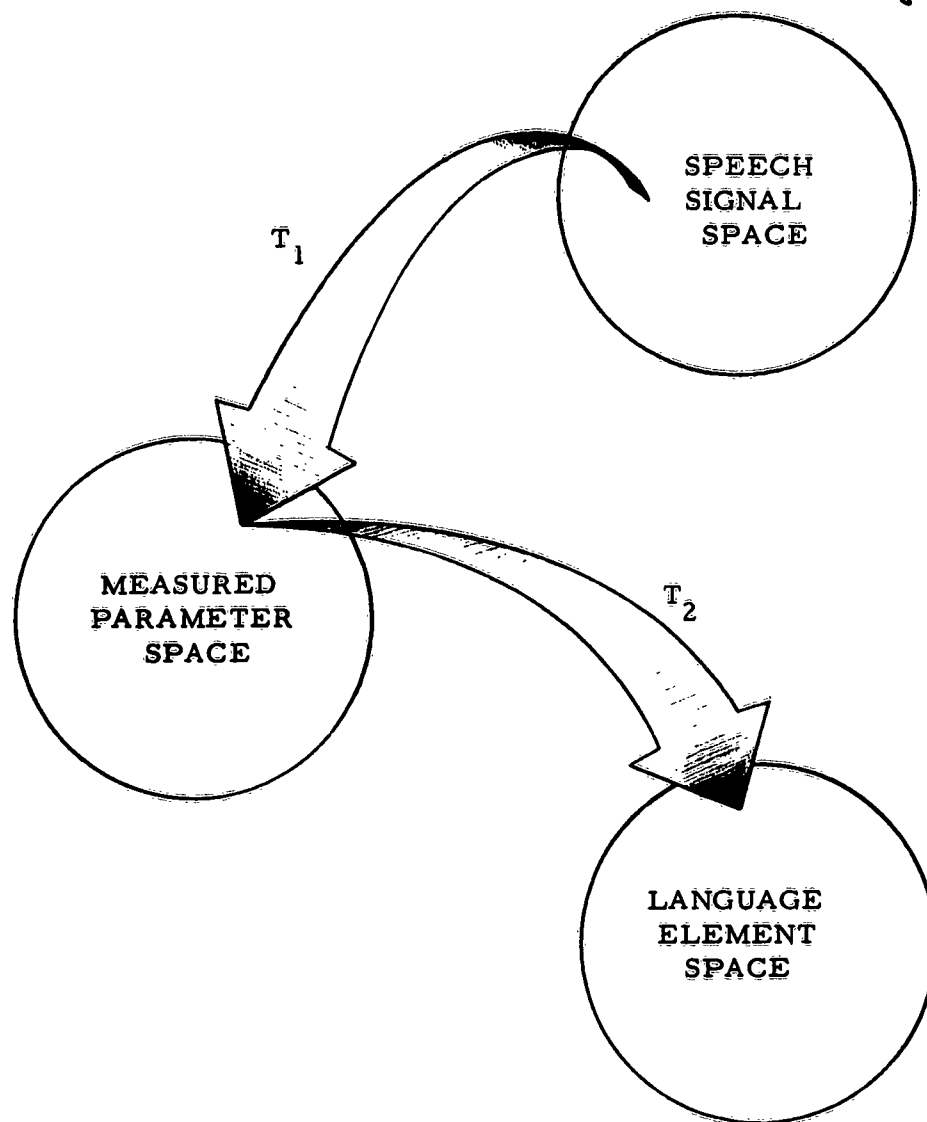


Figure 1. Two Basic Transformations in the Representation of Speech

The representation of speech in parameter spaces formed by two combinations of speech parameters is discussed in Section 3. Data are presented which indicate the storage requirements and accuracy associated with these representations.

Specific transcription and word recognition methods are described in Section 4. Results of experiments conducted to ascertain performance capabilities of these methods are also included in the fourth section, for one speech parameter space.

Conclusions regarding the type of speech representation which will prove most useful for preserving the information content of speech, and also serve as a convenient means of implementing automatic transcription techniques are presented in Section 5.

2. THEORETICAL BASIS FOR AUTOMATIC SPEECH TRANSCRIPTION

The transformation of speech into readable text involves two basic steps, or subsidiary transformations. As previously noted (Figure 1) the first involves association of a set of parameter values with each possible speech signal. The second involves the association of language elements with patterns of parameter values. Derivation of these two subsidiary transformations requires that a suitable list of parameters be selected, and a method be devised for associating patterns of these parameters with language elements. Also, a specific set of language elements must be selected. These three aspects of the speech transformation problem are discussed in the following subsections.

2.1 SELECTION OF LANGUAGE ELEMENTS

Several possibilities have been given consideration as language elements. The most frequently listed elements are words, syllables, phonemes, and phonetic elements, or "sounds". Several investigations have been conducted to determine the feasibility of using each of these language elements as a basis for transcription.* From the standpoint of ease of interpretation by a human reader, words are the most attractive language elements. However, with these elements the problem of selecting suitable parameters for representing speech is difficult to solve within reasonable limits on equipment complexity and/or storage requirements. Perhaps the basic reason for this is that the number of words required to represent a reasonably broad class of speech signals is large. A rudimentary vocabulary consists of several hundred words. This fact creates several obstacles to the construction of an automatic transcriber using words as the language elements. Notable among these is the difficulty of selecting parameters which are useful for separating more than a few words. Since words are composed of sequences of the intervals of speech corresponding to different states of the speech source, it is clear that parameters must be constructed in such a way as to produce different values for these sequences. This requirement suggests that parameters should be chosen by examining a given collection of words and selecting features of these specific words which tend to separate them. If the vocabulary is to remain the same, then this can produce a satisfactory result. If, however, the vocabulary is ever augmented, or even changed by replacement, then there is no guarantee that the selected parameters will produce a reasonable separation of the new words. Thus, from the standpoint of either restricting the speech which can be transformed satisfactorily or requiring major changes in the operations involved in speech representation in parameter space, words are unattractive as language elements.

*See for instance [7] for words, [16] for syllables, and [5] for phonetic elements or phonemes.

A related, practical difficulty which arises from the use of words as language elements is that the number of positions in parameter space which can conceivably correspond to words is large. If, for instance, periodic speech samples (spaced Δ seconds apart) are quantized in some manner with q possible different values, then the number of possible positions in parameter space is $q^{T/\Delta}$, where T is an indication of the duration of a spoken word. For most words, $\frac{T}{\Delta}$ is greater than 10, and if $q = 10^3$ (a conservative assumption) then parameter space may consist of as many as 10^{60} different points. Of course, if parameters are constructed from sequences of speech samples, then this number can be reduced tremendously. The selection of such parameters, however, very likely cannot be accomplished through the systematic examination of combinations of parameters as suggested in Section 2.1, because the number of different words and speech samples involved in a respectable vocabulary would be too large. Letting w denote the number of words in a vocabulary, and u denote the number of utterances of each word that would be used as a basis for learning the distribution of words in parameter space, suppose it is desired that all combinations of k parameters out of n candidates be examined. This would require that $\left(\frac{T}{\Delta}\right) (w) (u) \binom{n}{k}$ speech samples be processed. For $\frac{T}{\Delta} = 10$ samples per word, $u = 10$ utterances, $n = 10$ parameters, and $k = 8$, then 450,000 speech samples would have to be processed to obtain an indication of the distribution of only 100 words in the spaces formed by all combinations of the eight parameters. If 60 samples are obtained each second, then approximately 40 hours of speech would have to be processed to obtain the required data.

Another problem which arises from the use of words as language elements is that speech signals inherently must be segmented by some means into intervals of time corresponding to utterances of words. The transitions and other characteristics of signals, including silence intervals, apparently do not offer an unambiguous basis for performing this segmentation. This problem alone serves to restrict the use of words as basic language elements to the representation of words spoken in isolation. For continuous speech, most of the effort in recent years has been applied to the investigation of syllables or subsyllabic language elements.

If syllables are used as language elements, then some of the difficulties associated with the use of words are ameliorated. The correspondence between source states and syllables involves shorter sequences of intervals during which speech signals do not change significantly, and the number of different syllables required to represent a wide variety of speech is somewhat smaller* than the number of words in a comprehensive vocabulary. Unlike words, syllables need not (and probably cannot) be defined in a way which exactly corresponds to linguistic syllabification. One method which has been under study** for several years employs syllables defined as patterns of parameter values corresponding to utterances of standard, short words. In this system, the parameters consist of presence or absence of threshold crossings at the outputs of a filter bank, sampled at several different times. The samples are taken at times corresponding to significant changes in the speech signal. With 8 filters and 5 samples per syllable, the parameter space consists of 2^{40} possible patterns. Very likely, only a small percentage of these patterns would ever occur as the result of speech signals. In [6], for instance, it is suggested that ten to fifteen different patterns arise from a given syllable, and if 1000 syllables are required to adequately represent speech, then approximately 10^4 different patterns of parameter values would be used, assuming negligible overlap between syllables. This number places the use of syllables within the realm of practicality. The design of "Exact Match" devices for associating patterns of parameter values with syllables can exploit "either-or", "always-present", and "never present" conditions for each of the 40 binary parameters corresponding to a filter and sampling instant. For any single syllable, the "never present" condition will exist for most of the parameters, thus allowing for construction of a relay "tree", consisting of only a few relays, for recognition of each syllable.

As remarked above, the two primary ways in which the use of syllables constitutes an improvement over the use of words as language elements are (a) the number of significant changes which occur in speech signals during intervals corresponding to language elements is reduced, and (b) the number of different language elements needed to adequately represent speech is reduced. These changes permit the use of a smaller parameter space, and simplify language element recognition (1000 syllables instead of perhaps 5000 words for a comprehensive vocabulary). To some investigators, it appears that the use of sub-syllabic language elements would offer even greater simplification of the speech recognition problem

*It is suggested in [6] that 1000 syllables would suffice to accurately represent an unrestricted vocabulary.

**[6].

by the same means: reduction in the size of parameter space, and reduction in the number of language elements (i. e., the number of alternatives to which each pattern of parameter values must be assigned). Representation of speech with phonemes, for instance, has been the goal of several investigations. As with syllables, the definition of phonemes for the purpose of automatic speech transcription necessarily differs from the linguistic definition.* For automatic speech transcription, a phoneme consists of those patterns of parameter values which result from utterances judged by either a human or other means to be a distinctive speech sound. If the judgment is made by a human, then these language elements comprise a phonetic alphabet. From the standpoint of ease of interpretation by a human reader, a phonetic alphabet evidently would be quite satisfactory. Although the reader would be required to learn the alphabet, this can be accomplished quite easily.**

In view of the fact that words and syllables are composed of sequences of phonetically distinguishable intervals of speech, it might be expected that speech signals will change less during intervals assigned to symbols in a phonetic alphabet than during intervals which would be assigned to syllables or words. Thus, it is possible that a smaller parameter space may suffice to distinguish between phonetic speech elements than is required for the longer elements. However, it has been contended that no matter what parameters are used, the variations in manifestations of different speech sounds (in different environments, from different speakers, etc.) in parameter space are so large that separation of these sounds is not possible. The question of feasibility of separation of speech sounds (in a parameter space) raised by these contrary points of view probably can be resolved in the affirmative only by demonstration, i. e., by developing operations which actually produce different outputs corresponding to different sounds.

If such a parameter space can be found, then the use of phonetic elements to represent speech vastly simplifies the problem of associating patterns of parameter values with the language elements. Approximately 40 phonemes are considered sufficient to adequately represent speech. Thus, the number of alternatives for assignment of a pattern of parameter values is only a few dozen, compared with a thousand or more, as would be required for adequate representation with syllables or words.

*According to [4], "a phoneme is the minimum feature of the expression system of a spoken language by which one thing that may be said is distinguished from any other thing which might have been said".

**It has been suggested that a phonetic alphabet facilitates reading, and has been adopted for use in a few schools. See, for instance, [12].

In keeping with examining simple methods first, we have directed our attention on this project to the use of phonetic elements. As will be shown in Section 3, enough separation between some of these elements can be achieved with a minimal parameter set, to indicate that addition of other parameters will provide essentially non-overlapping patterns of parameter values corresponding to different speech sounds.

Rather than dwell on the distinctions between linguistic and operationally defined phonemes, we have somewhat arbitrarily set up a phonetic alphabet which consists of symbols corresponding to intervals of speech during which very little change can be detected acoustically. These symbols, and examples of words whose normal pronunciation produces speech sounds corresponding to these symbols, are listed in Table 1. Also indicated are phonemes whose utterances produce the speech sounds. The phonetic elements are labeled in an arbitrary but suggestive way which allows for convenient print-out from the general purpose digital computer with which transcription methods are simulated.

2.2 SELECTION OF SPEECH PARAMETERS

The problem of extracting clues from a speech signal which contain sufficient information to identify the language elements being uttered can be formulated and attacked in two somewhat different ways. One approach to the problem consists of drawing up a list of features of speech which are phonetically distinguishable by humans and which it is believed will serve to classify speech signals into sequences of phonetic language elements. These features generally correspond to different states of the human speech source, i. e., articulatory states of the vocal tract -- for example, the vocal cord vibration rate, the mouth opening, and positions of the tongue and lips. Since the correspondence between source states and the generation of phonetic language elements is relatively well known, the representation of speech as a sequence of these language elements can readily be solved if a means can be devised to measure automatically the phonetic, or "distinctive"*, features of speech.

An example of such a list of distinctive features is shown in Table 2. ** As indicated in the table, determination of the presence or absence of ten speech features is evidently sufficient for a human to distinguish between 35 different phonemes (which could be used to represent English quite adequately).

*[1].
**[5].

TABLE 1 - PHONETIC ALPHABET

SOUND				EXAMPLES	SOUND GROUP CHARACTERISTICS
Group	Number	Designation			
		Recomp	IPA Phoneme(s)		
I	1	AW	ɔ	<u>A</u> LL	VOWELS
	2	OO	u	<u>P</u> OO <u>L</u> , <u>W</u> AI <u>L</u>	
	3	U	U	<u>P</u> U <u>L</u> L	
	4	UR	ɜː	<u>B</u> I <u>R</u> D, <u>M</u> A <u>K</u> E <u>R</u>	
	5	AH	ɑː, ɒ	<u>F</u> A <u>T</u> H <u>E</u> R, <u>O</u> DD	
	6	UH	ʌ, ɹ	<u>S</u> U <u>N</u> , <u>S</u> O <u>F</u> A	
	7	O	o, ou	<u>N</u> O <u>T</u> A <u>T</u> I <u>O</u> N, <u>G</u> O	
	8	A	a, æ	<u>A</u> S <u>K</u> , <u>S</u> A <u>T</u>	
	9	EH	e	<u>S</u> E <u>T</u>	
	10	I	i	<u>S</u> I <u>T</u>	
	11	EE	i, j	<u>B</u> E <u>E</u> T, <u>Y</u> O <u>U</u>	
II	12	L	l	<u>L</u> U <u>L</u> L	LIQUIDS
	13	R	r	<u>R</u> E <u>A</u> R	
	14	W	w	<u>W</u> A <u>I</u> L	
III	15	M	m	<u>M</u> A <u>I</u> M	NASAL
	16	N	n	<u>N</u> O <u>O</u> N	CONSONANTS
	17	NG	ŋ	<u>S</u> I <u>N</u> G	
IV	18	B	b	<u>B</u> I <u>B</u>	VOICED
	19	D	d	<u>D</u> E <u>E</u> D	STOP
	20	G	g	<u>G</u> I <u>V</u> E	CONSONANTS

TABLE 1 (Cont.)

SOUND				EXAMPLES	SOUND GROUP CHARACTERISTICS
Group	Number	Designation			
		Recomp	IPA Phoneme(s)		
V	21	Z	z	<u>Z</u> ONE	VOICED FRICATIVE CONSONANTS
	22	V	v	<u>V</u> AL <u>V</u> E	
	23	TJ	ʒ	E <u>I</u> TH <u>E</u> R	
	24	ZH	ʒ	VI <u>S</u> ION	
VI	25	T	t	<u>T</u> OO <u>T</u>	UNVOICED
	26	P	p	<u>P</u> EE <u>P</u>	STOP
	27	K	k	<u>C</u> A <u>K</u> E	CONSONANTS
VII	28	H	h	<u>H</u> A <u>I</u> L	UNVOICED
	29	WH	hw	<u>W</u> H <u>A</u> LE	
VIII	30	F	f	<u>F</u> I <u>F</u> E	FRICATIVE
	31	TH	θ	<u>T</u> H <u>I</u> N	
IX	32	S	s	<u>C</u> E <u>A</u> SE	CONSONANTS
	33	SH	ʃ	MI <u>S</u> S <u>I</u> ON	
X	34	CH	tʃ	<u>C</u> H <u>U</u> R <u>C</u> H	AFFRICATES
	35	DJ	dʒ	<u>J</u> U <u>D</u> GE	

The basic difficulty with this approach arises from the need to develop operations which can be performed on the speech waveform to determine the presence or absence of the specified distinctive features. Although study of the mechanisms by which the speech source generates language elements has yielded considerable knowledge of speech waveform characteristics, notably energy distributions in time and frequency, no reliable correspondence between such measurable characteristics and the presence or absence of distinctive features (as judged by humans) has as yet been developed. If this approach to the speech processing problem is pursued vigorously, then major emphasis is inevitably placed on attempts to develop better ways to determine presence or absence of the distinctive features.

Although this approach recognizes the basic problem of representing speech in terms of measurable parameters, it tends to deify certain pre-selected parameters as those which should be used to classify language elements. Unfortunately, mechanizations of the judgment of parameter values (i. e., presence or absence of distinctive features) have generally proved unsatisfactory in one way or another.

The other general approach to the speech processing problem differs from the first primarily in the way in which parameters are selected. First, parameters are considered to be defined only in terms of operations performed on the speech waveform. Although considerable guidance in the selection of suitable operations for distinguishing speech sounds is provided by knowledge of the manner in which humans solve the problem, no pre-determined list of speech characteristics or distinctive features is drawn up as an unalterable goal. Instead, several candidates for useful parameters are selected not only on the basis of their possible potential for classifying language elements, but also on the basis of ease of implementation. These parameters are examined to determine which language elements can be classified through their measurement. By study of combinations of parameters it is possible to pinpoint the language element confusions which remain to be resolved, as well as the combination of parameters which achieves the greatest language element separation. Generally, it is expected that a close examination of the parameter values associated with language elements not distinguished with the initial set of parameters will yield suggestions for other operations or parameters which will serve to classify these elements. By systematically introducing new parameters for the purpose of resolving these remaining language element ambiguities, it is anticipated that an implementable set of speech parameters will be obtained which is sufficient to classify language elements. We have pursued this course using phonetic elements, or speech sounds, as the language elements.*

*See Table 1 in Section 2. 1.

PHONEMES

DISTINCTIVE FEATURES

	i	I	u	U	ε	æ	o	ɔ	A	ʌ	r	l	w	j	m	n	ŋ	f	s	θ	ð	ʒ	ʃ	z	ʒ	ʃ	p	t	k	b	d	g	h
1. Sonorant/Non-Sonorant	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2. Consonant/Non-Consonant	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-
3. Compact/Non-Compact																																	+
4. Diffuse/Non-Diffuse	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
5. Grave/Acute	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-
6. Tense/Lax	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-
7. Flat/Plain																																	
8. Nasal/Oral																	+																
9. Continuant/Interrupted																		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-
10. Strident/Mellow																		+															-

TABLE 2. Correspondence Between Speech Features and Phonemes*

* From [5]

It should be emphasized at the outset that from the standpoint of attaining or improving a speech transcription capability, speech parameters should be selected on the basis of the degree to which different combinations of parameter values are obtained for utterances of different sounds. This does not mean that different parameter values will result for different sounds, when parameters are considered individually. To illustrate this point, consider the diagram in Figure 2. As depicted in the diagram, two sounds may result in values of two parameters which are quite similar, and involve considerable overlap between the two sounds, when either parameter is measured alone. However, simultaneous measurement of the two parameters may produce a non-overlapping distribution of the two sounds in the 2-dimensional parameter space, as shown in the diagram. This rather elementary observation suggests that (a) the systematic introduction of new parameters, as outlined above, will produce an efficient speech representation in terms of storage requirements, and (b) essentially rules out the selection of speech parameters solely on the basis of separation of single parameter values arising from different sounds.

In this study we have undertaken initially to develop a set of speech parameters which serve to distinguish primarily between voiced sounds. The speech mechanism for voiced sounds may be thought of as an acoustic pulse generator (the vocal cords) exciting a multiply resonant cavity (the vocal tract, including nasal cavities). The vibration rate of the vocal cords is commonly associated with the pitch frequency. The several resonances, each of approximately 90 cps in bandwidth, are several times higher in frequency than the fundamental and vary considerably from sound to sound and exhibit some variation from speaker to speaker. These resonances of the vocal tract, called formants, give rise to local peaks in the energy spectra of speech samples. The location of these peaks may be regarded as indications of formant positions for voiced sounds. It has long been recognized that formant characteristics serve to distinguish fairly well between the vowel sounds, and also carry considerable information on other voiced sounds. We have therefore chosen to use an estimate of formant positions (location of spectrum peaks) as the initial set of parameters. The automatic extraction of local spectrum peaks can be accomplished with a spectrum analyzer of the type commonly used in vocoders.

A second set of parameters has been selected to obtain information on the shape of the energy density spectra of speech signals. Study of "sections" of speech (i. e., energy density spectra) on an audio signal analyzer, or observation of commutated samples of vocoder channel outputs displayed on an oscilloscope, provides a strong indication that different

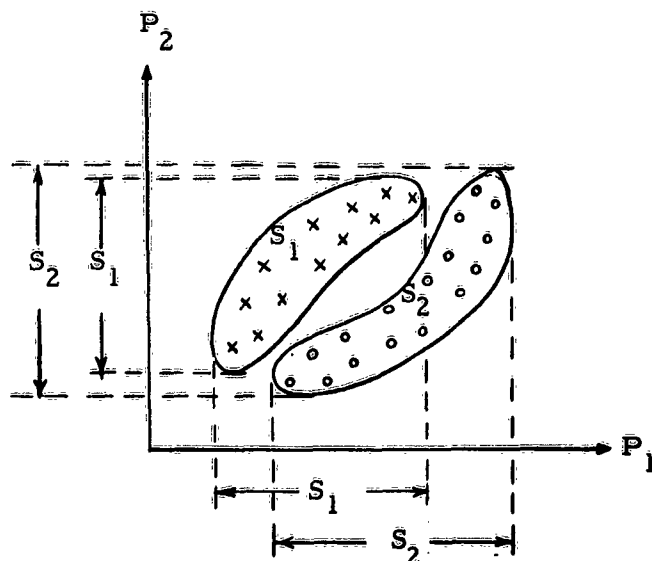


Figure 2. Hypothetical Distribution of Two Speech Sounds in a Two-Dimensional Parameter Space

sounds give rise to significant differences in spectral shapes. Various possibilities exist for obtaining spectral shape information. Rather than develop a number of operations, each of which would be designed to detect spectrum shapes associated with a few sounds, we have investigated initially a set of parameters which is not only expected to provide useful information on all sounds (including unvoiced sounds), but also is easily instrumented with a minimum of adjustment.

The two characteristics of any, possible unknown, function of a quantity x (say, $f(x)$), which have come to be regarded as perhaps the most important for characterizing the shape of $f(x)$, are the first moment about the origin and the second moment about the mean. Considering the energy density spectrum of a speech sample as a function of frequency $S(f)$, these quantities can be defined for speech by

$$\mu = \frac{M_1}{M_0} = \text{spectrum mean}$$

$$\sigma = \left\{ \frac{M_2}{M_0} - \mu^2 \right\}^{1/2} = \text{spectrum spread}$$

where $M_0 \equiv \int S(f) df = \text{spectrum area, or spectrum zero-th moment}$

$$M_1 \equiv \int f S(f) df = \text{spectrum first moment}$$

and

$$M_2 \equiv \int f^2 S(f) df = \text{spectrum second moment.}$$

It is expected that different values of μ and σ will result from utterances of different speech sounds. To obtain μ and σ it is sufficient to measure M_0 , M_1 , and M_2 - each of which can be obtained through linear operations on the outputs of a vocoder spectrum analyzer. While these quantities may not be sufficient to differentiate between all sounds, it is anticipated that their measurement will afford a significant improvement in transcription capability over that attainable with spectral peaks alone, for unvoiced as well as voiced sounds.

A third set of parameters, which would serve to advance transcription capability significantly, consists of measurements to ascertain those times at which parameters of the type described above are undergoing relatively rapid changes. These changes would serve to segment speech into intervals corresponding to either the language elements being used to represent speech, or some other elements from which the language elements can be ascertained. By processing the information obtained in such a segment of speech before rendering a decision as to which sound is being uttered, the reliability of decisions can be improved over the performance attainable through decisions rendered more frequently.

Assuming that a set of measurements, or operations, to be performed on a speech signal have been formulated, consider now the behavior of speech as represented in parameter space. This space can be formed by considering each parameter to be a coordinate direction in (for convenience in visualization) a rectangular coordinate space. A speech signal, $s(t)$, could then be represented in vector form.

$$s(t) \longrightarrow \underline{v}(t) = [v_1(t), v_2(t), \dots, v_n(t)],$$

where $v_i(t)$ is the time-varying result of the operation on the speech signal defined by the i -th parameter; i. e., $\underline{v}(t)$ indicates the point in the n -dimensional parameter space (formed by the n parameters v_1, v_2, \dots, v_n) into which the speech signal is mapped, at the time instant, t . As speech is uttered, the point \underline{v} moves about in parameter space in some manner corresponding to the sequence of sounds being uttered. If a good set of parameters has been selected, then the point \underline{v} will lie within different regions of parameter space during intervals corresponding to different sounds. Solution of the speech transcription problem requires that such a set of good parameters be found, and an easily implemented method be devised for describing the regions in parameter space corresponding to the different sounds.

As a first step towards simplification of equipment, it is possible to quantize speech into intervals of time during which the point \underline{v} changes insignificantly. Since a speech signal envelope has a bandwidth of about 25 cps, no significant information loss is suffered if the envelope detected outputs of a vocoder spectrum analyzer are sampled periodically, at a rate of 50 samples per second or higher.* Further, any parameters which

*This assertion has been verified through listening tests with speech synthesizers utilizing periodic speech samples as inputs. Speaker fidelity is also preserved.

are obtained through linear operations on the speech signal envelope may also be sampled at the same rate with essentially no loss in information. Thus, a speech signal, $s(t)$, may be represented as a sequence of positions in parameter space, described in vector form by

$$s(t) \Rightarrow \underline{v}_i = [v_{1i}, v_{2i}, \dots, v_{ni}], \quad i = 1, 2, \dots,$$

where the subscript "i" indicates the position in parameter space occupied at the i-th sample instant. The time separation of adjacent samples, Δ , must be no greater than approximately 20 milliseconds.

As a second step toward simplification of equipment, it will be desirable to reduce the number of possible positions which can be occupied in parameter space by quantization of the speech parameters. However, in quantizing parameter values, considerable care must be exercised to avoid the creation of ambiguities in parameter space which are large with respect to the separation of different sounds (as represented in parameter space). This problem deserves as much attention as the selection of parameters, since quantization itself is one of the operations which defines a parameter.

The potential of several operations constituting parameters of the type suggested above for distinguishing between vowel sounds is examined in Section 3. The critical importance of proper quantization of parameter values is demonstrated with one of these parameters.

2.3 PATTERN RECOGNITION METHODS

The selection of speech parameters which take on different values during intervals of speech corresponding to utterances of different sounds constitutes the first, and most important step toward achieving an automatic transcription capability. Once a set of parameters has been selected which serve to separate speech sounds in parameter space, the problem of associating patterns of parameter values with language elements arises. This problem consists of two parts. First, the distribution of speech sounds in parameter space must be ascertained; i. e., the patterns of parameter values which arise from a speech sound (and preferably their relative frequency of occurrence) must be found for all sounds in the phonetic alphabet. Second, a satisfactory means of partitioning parameter space into regions corresponding to the different speech sounds must be devised. If it can be ascertained that there exist no points (i. e., patterns) in parameter space which ever arise from utterances of more than one sound, then this problem is trivial. Implementation of the decision boundaries requires only that the equivalent to a table look-up operation be implemented.

We shall refer to these two parts of the problem of associating points in parameter space with speech sounds as (1) finding the distribution of sounds, and (2) establishing decision boundaries.

The complicating feature of the problem of finding the distribution of classes of events (in this case, sounds) as represented in a parameter space, is that in most practical situations the classes are known only through a finite set of sample events. The number of possible patterns of parameter values usually exceeds by far the number of sample events which can be obtained. Thus, solution of the first part of the pattern recognition problem requires that some doctrine be applied to decide whether any of the parameter patterns which have not occurred in the sample events should be regarded as belonging to any of the classes, and if so, to which classes. It must also be decided whether any of the sample patterns arising from events belonging to one class might also ever occur as manifestations of events belonging to another class. Since the number of available sample events is usually small, one is faced with the necessity for constructing some conception of the distribution of classes in parameter space, based on incomplete information concerning the class association of a sparse collection of sample points. If the chosen parameters

produce tight, widely separated clusters of points in parameter space corresponding to the different classes, then a few samples should suffice to "learn" the distribution of classes sufficiently well to avoid incorrect associations of points with classes. However, if the parameters do not produce such a situation, then either many samples must be obtained to learn the nature of the distribution of classes, or if for some reason this is not possible, a possibly hazardous estimate of the distribution must be made on the basis of data at hand.

Perhaps the most complete description of the distribution of speech sounds in parameter space which one can ever hope to obtain is the probability density functions of points in parameter space, \underline{y} , conditioned on each of the N speech sounds, S_i , $i = 1, 2, \dots, N$. If parameter space consists of discrete points (as in the case when all parameter values are quantized), then the probability density function of \underline{y} , conditioned on the i -th speech sound, $p_i(\underline{y} | S_i)$, is equivalent to the probability, $P_i(\underline{y} | S_i)$, that the point \underline{y} will occur in parameter space when the i -th speech sound is uttered. The fact that a restricted number of sample patterns may be available from which the distribution of sounds in parameter space can be inferred, has motivated the development of a variety of methods of estimating the nature, or particular characteristics, of the functions $P_i(\underline{y} | S_i)$, using a limited amount of data.

Once some conception of the distribution of speech sounds in parameter space is obtained, the problem of partitioning the space into non-overlapping regions corresponding to speech sounds can be attacked. A method of establishing decision boundaries which has come to be regarded as an optimum method consists of calculating the likelihood that a given point, \underline{y} , has arisen from the i -th sound, S_i , $i = 1, 2, \dots, N$, and choosing the sound for which this quantity is highest. If the a priori probability of occurrence of a speech sound is assumed to be the same for all sounds, then this Maximum Likelihood method is equivalent to the establishment of decision boundaries according to the following rule:

If $P_i(\underline{y} | S_i) \geq P_j(\underline{y} | S_j)$ for $i = 1, 2, \dots, N$, then associate \underline{y} with the i -th sound.

Thus, the maximum likelihood method requires only the comparison of values of the functions $\{P_i(\underline{y} | S_i)\}$, which have already been established as goals for the description of speech sounds as represented in parameter space.

A direct approach to the problem of estimating the functions $\{P_i(\underline{y} | S_i)\}$, consists of constructing histograms over parameter space from a large number of independent samples of each of the speech sounds. If one is able to obtain enough* samples, there appears to be no better way to proceed.

Most of the pattern recognition methods which have been developed represent attempts to exploit some a priori notions of the functions, $\{P_i(\underline{y} | S_i)\}$ in the estimation of certain features of these functions. For instance, it may be assumed that each of these functions is unimodal, i. e., possesses a single local maximum. Under this assumption the use of multiple order linear discriminant functions may lead to good results. This method involves the use of hyperplanes of the form $y_i = (\underline{a}_i \cdot \underline{y}) =$

$$\sum_{k=1}^n a_{ik} v_k \quad \text{for the decision boundaries, where the coefficients, } \underline{a}_i =$$

$(a_{i1}, a_{i2}, \dots, a_{in})$, can be determined in a variety of ways. Although several hyperplanes can be used to bound each class (speech sound) from each of the other classes, it has rarely been suggested that any more than a single hyperplane for each pair of classes be used since even this number becomes intolerably large when the number of classes is greater than a dozen or so. The rather great reliance which has been placed on the use of linear discriminant functions for establishing decision boundaries in classification problems, suggests that potential hazards associated with their use have not been fully appreciated. If any of the functions $\{P_i(\underline{y} | S_i)\}$ are not unimodal, then a hyperplane may be completely ineffective in partitioning parameter space into regions corresponding to different classes. Consider for instance the hypothetical distribution of two sounds in a two-dimensional parameter space as depicted in Figure 3. Although these two sounds are non-overlapping and even tightly clustered (in multiple modes) and widely separated, there exists no straight line which can be drawn to completely separate the two classes.

Another popular method consists of using a single linear form for each class. This method consists of correlating the pattern, \underline{y} , with a single representative of the i -th class, \underline{b}_i , and choosing the class for which the correlation is highest, after normalization:

*An attendant problem is that an implementable, general criterion by which the number of available samples can be judged "enough" or not, has not yet emerged from the large amount of study which has been directed to the question over the years.

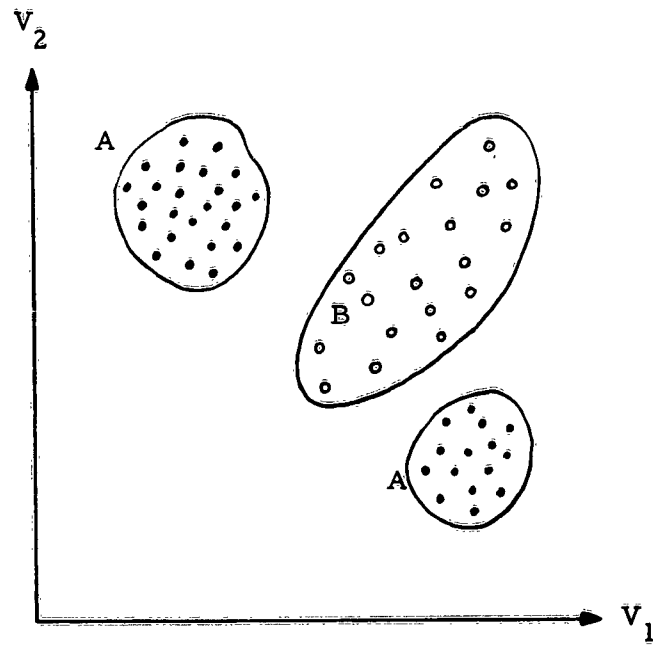


Figure 3. Hypothetical Multimodal Distribution of Two Speech Sounds in a Two-Dimensional Parameter Space

Choose the j-th class if

$$\frac{(\underline{b}_j \cdot \underline{v})}{\|\underline{b}_j\| \|\underline{v}\|} \geq \frac{(\underline{b}_i \cdot \underline{v})}{\|\underline{b}_i\| \|\underline{v}\|}$$

for all i. Again, a variety of methods can be applied to determine suitable class representatives, \underline{b}_i . The most common choice is the sample mean vector of each class. This method provides the same decision boundaries as would be obtained with the Maximum Likelihood method, if all the classes have symmetric Gaussian distributions, with equal variances in the parameter space. If the classes do not possess such distributions, then this method may or may not produce good results.

Correlation with a stored reference constitutes an example of the treatment of pattern recognition as a two-step problem wherein (1) a representative pattern is selected for each class, and (2) sample patterns are associated with classes according to which representative is "closest" to the sample, where the "distance" between two vectors is defined in some way. The attributes of a variety of methods of measuring distance have been studied extensively*, with the result that as constraints are removed from the form to which the measure of distance is limited, better recognition capability is achieved.

A related approach** consists of establishing nonlinear decision boundaries in parameter space, where coefficients of second, third, and higher powers of parameter values (in contrast with coefficients of linear terms) are selected as a basis for fitting complicated shapes around regions associated with different classes. Again, a variety of criteria can be applied to select values for the coefficients. This approach provides much greater potential for separating multi-modal distributions than linear methods, but as a general rule, more samples are also required to obtain an accurate placement of the nonlinear boundaries.

There are, of course, many different pattern recognition techniques which have been developed in the past few years, some of which offer computational simplicity in lieu of potential accuracy, and vice versa. Although the techniques mentioned here encompass many of the methods which have

*[8], [10].
 **[8].

been developed, there are many others which either exist now or which will no doubt come along in the future. The question of which method would be best for recognizing speech sounds can only be answered by either (a) obtaining a large number of samples of each sound from which a good estimate of the functions $\{P_i(\underline{v} | S_i)\}$ can be obtained, or (b) using each method and comparing the results. In the course of this study, we have stressed the development of techniques by which sufficient data can be collected to estimate the functions $\{P_i(\underline{v} | S_i)\}$.

Two further points should be emphasized. First, if speech parameters can be found which produce no overlap between speech sounds, and if the number of different combinations of parameter values is small, then there is nothing to be gained in using any special pattern recognition method such as setting up a particular type of discriminant function. It would be sufficient to compare a speech sample with a "reference library" of patterns and associate the sample with that sound to which the reference duplicate of the sample (if one exists) corresponds. If an exact match does not occur between the speech sample and some member of the reference library, then a variety of possibilities exist. For instance, the sound corresponding to the last match could be assumed to persist. Or if more than one speech sample is obtained for a sound (as is the case for most speech sounds) then the decision could be deferred.

The implementation of such a method can be extremely simple if the number of patterns which actually arise from speech is not unreasonably large. Although it has been contended* that speech sounds produce too many manifestations in a parameter space to allow this method to be employed, no proof of this contention has been provided. To do so it would be necessary to show that all possible parameter spaces would produce wide variations, since the variation within each speech sound as represented in parameter space is determined by the parameters themselves. On the other hand, the fact that intelligible speech is produced by speech synthesizers operating on parametric representations of speech indicates that parameters may exist which are relatively invariant for a single speech sound, and yet produce different values for different sounds.

*For instance in [6].

The other point is that if more than one decision is made during an interval corresponding to a single speech sound, then it is not necessary that a match be obtained for every speech sample, in order to use the Exact Match method outlined above. Thus, it may be possible to reduce the reference library to a relatively small number of patterns of parameter values, without sacrificing the accuracy with which speech sounds can be identified. In the next section, results of a few experiments are presented as a basis for an initial estimate of the degree to which a given size reference library can be expected to cover (i. e., match) all possible samples of speech sounds.

3. SPEECH REPRESENTATION IN PARAMETER SPACE

As discussed in Section 2.2, it has been anticipated that positions of local maxima in sample speech energy density spectra will provide sufficient information to distinguish fairly well between the vowel sounds, and will also serve as useful clues for identifying other voiced sounds. Also, it is highly likely that vowel sounds cannot be recognized adequately if some indication of formant positions is not available. Therefore, we have considered spectral peak patterns to constitute a minimum parameter set for voiced sounds. The ways in which these estimates of the formant positions and other speech parameters have been extracted in this investigation are described in the first of the following two subsections. In the second subsection, the results of an investigation to determine the distribution of sounds in the resulting parameter spaces are reported for two combinations of parameters.

3.1 SPEECH PARAMETER EXTRACTION

Although intuitive conceptions of speech parameters can be described in terms that are readily accepted and understood by everyone, the problem of extracting numerical values of speech parameters in a way which will be deemed satisfactory by even a few people is still a difficult one to solve. For instance, to ascertain whether a particular method of extracting, i. e., estimating, formant positions is satisfactory or not, it can be contended that measurements of the vocal cavities in the speech source must be recorded simultaneously and compared with the numerical values obtained for the estimates of formant positions. However, as pointed out in Section 2.2, it is not absolutely necessary that the question of how well a parameter is being measured ever be raised. If one adopts the point of view that a parameter can be defined precisely only in terms of the operations actually performed on the speech waveform to obtain parameter values, then perforce the parameter is always extracted properly. The appropriate question then becomes: "What are the operations to be performed on the speech waveform (i. e., parameters) to obtain different numerical values for the different speech sounds?".

To answer this question, the quantities listed in Table 3 have been chosen as initial candidates for suitable parameters to facilitate speech transcription. In the course of this study, methods of extraction have been devised and applied for each of these quantities. The specific operations performed on the speech waveform to obtain numerical values of these quantities are described in the following paragraphs.

TABLE 3 - LIST OF SPEECH PARAMETERS INVESTIGATED

<u>Description</u>	<u>Notation</u>
Location of Peaks in Speech Sample Energy Density Spectrum (quantized into 18 frequency channels)-- indication of formants	$\mathcal{P} = (p_1, p_2, \dots, p_{18})$
Ratio of outputs of high pass and bandpass filters-- voicing indication	V
Area under Speech Sample Energy Density Spectrum-- Speech Sample Signal Energy	M_0
First Moment of Speech Sample Energy Density Spectrum-- Spectrum Mean	M_1
Second Moment of Speech Sample Energy Density Spectrum-- with the first moment and spectrum area, a measure of Spectrum Spread	M_2
Input Speech Envelope Amplitude	E
Normalized Speech Envelope Amplitude	E_0
Sum of Magnitudes of Forward Differences of Samples of the above parameters--speech segment boundaries	ΔS

The block diagram in Figure 4 indicates the major processing steps involved in the approach being taken to speech transcription in this study. These are extraction and periodic sampling of speech parameters such as listed in Table 3, further operation on and quantization of these samples, and association of patterns of the resulting quantized parameters with phonetic elements.

3.1.1 Description of Equipment

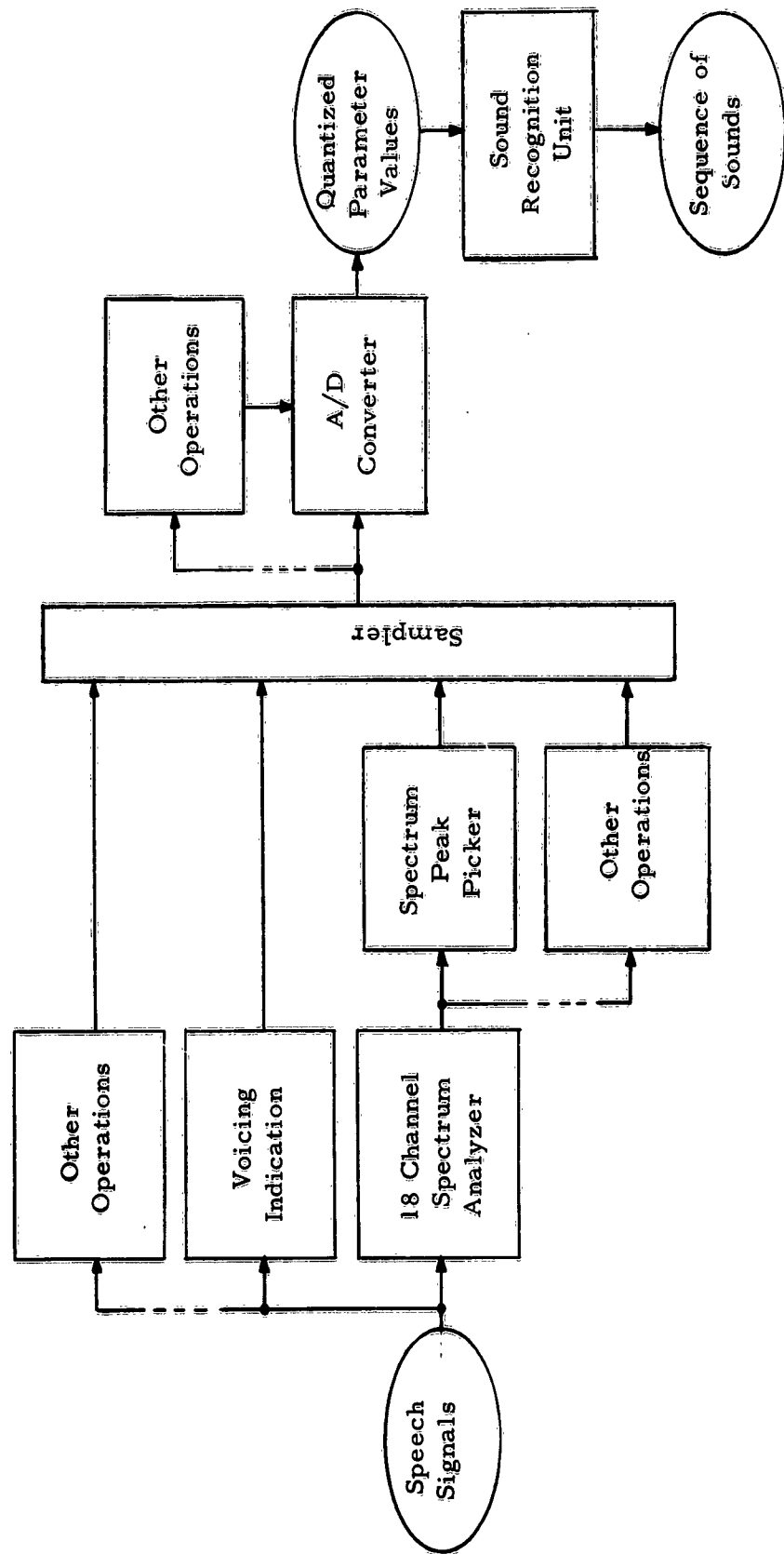
To obtain data as well as demonstrate the feasibility of this approach to speech transcription, laboratory equipment has been constructed in the Communication Sciences Laboratory for use on this and other speech processing projects. The configuration of this equipment for the parameters listed in Table 3 is shown in Figure 5. The three functions performed are (1) speech signal conditioning, (2) parameter extraction, and (3) data format conversion.

For signal conditioning, the first operation performed on a speech signal is pre-emphasis of the high frequencies. The pre-emphasis network serves to accentuate the relatively weak higher formants of voiced speech signals, and the sometimes extremely low-level unvoiced speech sounds. Although several adjustments of the pre-emphasis network were made during the course of this study, its final characteristic consists of approximately 6 db gain per octave above on kcps. The envelope detected output of this network, E , was studied as a possible candidate for a normalizing parameter for others.

After pre-emphasis, the speech signal is next passed through an AGC network. This network constrains the output to less than 6 db variation for a 20db range in the input signal level. The envelope detected output of this network, E_o , was also studied as a possible normalizing factor for other parameters.

After AGC, the signal is passed to an 18-channel parallel filter bank. The characteristics of these filters are indicated in Figure 6.

In the current experimental setup, the 18-channel filter bank output, denoted $\underline{f} = (f_1, f_2, \dots, f_{18})$, is fed to five parameter extractors: the peak picker, the zero-th, first and second moment calculators, and the speech



W/F 62-186

Figure 4. Block Diagram of Approach to Parameter Extraction and Speech Signal Mapping

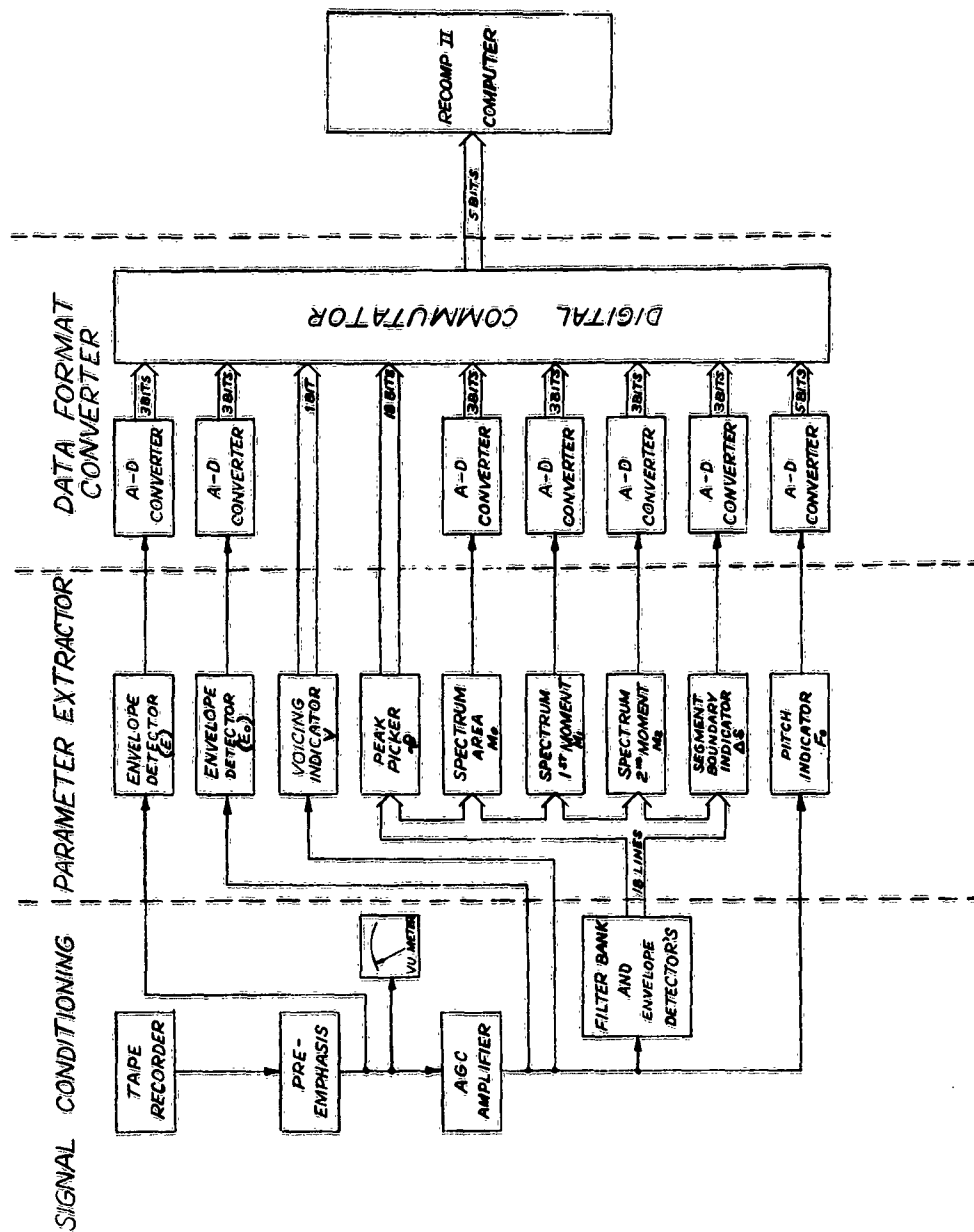


Figure 5 Block Diagram of Experimental Speech Processing Equipment

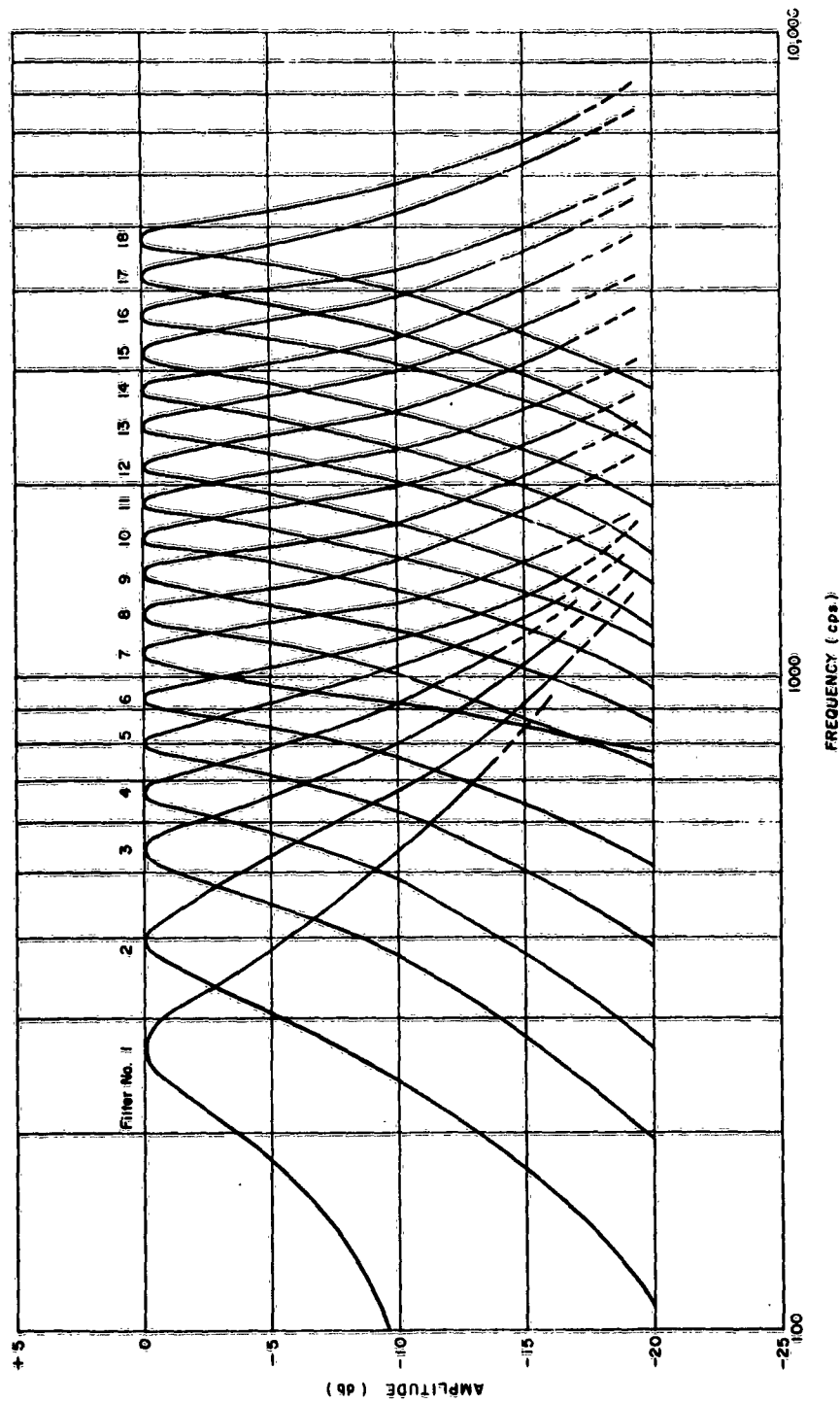


Figure 6 Litton Vocoder Spectrum Analyzer Filter Bank Response

segment boundary indicator. The peak picker is a unit designed to locate and identify those channels in which the energy density spectrum of the past few milliseconds of speech possesses a local maximum. This unit operates as indicated in Figure 7. The envelope detected output, x_n , of the n -th filter (channel) is fed to two adjacent comparators (C), one of which is preceded by a circuit (ϕ^+) which passes the greatest of two inputs. One of these two inputs is x_n , and the other input is a constant, K . The ϕ^+ circuits serve to prevent the indication of a peak in a channel unless the peak amplitude is greater than the adjustable threshold, K . If both comparators produce outputs indicating that the quantity x_n is larger than x_{n+1} and x_{n-1} , then a peak is indicated at the output by routing the adjacent comparator outputs to an AND gate corresponding to the n -th channel.

In view of their central role in the identification of voiced sounds, it is of interest to note the degree to which the peak patterns (as obtained with Litton's 18 channel vocoder and peak picking unit) correspond to other methods of extracting indications of formant positions. Although a thorough investigation has not been undertaken as yet, a few comparisons with a conventional method of extracting formant position estimates by hand show that the peak picking method produces quite similar results. Specifically, a method of measuring (i. e., estimating) formant positions which has been employed for years involves tracing (by hand) indications of local energy peaks as observed in spectrograms. This method can be compared with the automatic peak picking method by simply quantizing the tracings into the same 18 frequency channels employed in the peak picker, using the same speech sample for each method.

The result of such a comparison is shown in Figure 8 for the word "ONE". Although there are differences in the two estimates of formant positions, it is not clear which of these methods provides a more accurate indication. In order to assess accurately the quality of either method, it would be necessary to make careful measurements on the speech source, as well as detailed recordings of the speech source states (mouth opening, tongue position, etc.) during the utterance. It can be concluded from an examination of several such comparisons between the two methods of estimating formant positions, that the peak picking method tends to produce formant indications which resemble very closely those obtained by hand.

The moment extractors implement the calculation of zero-th, first and second moments of the speech energy density spectrum. In terms of the filter bank outputs, $\underline{f} = (f_1, f_2, \dots, f_{18})$, these parameters can be written

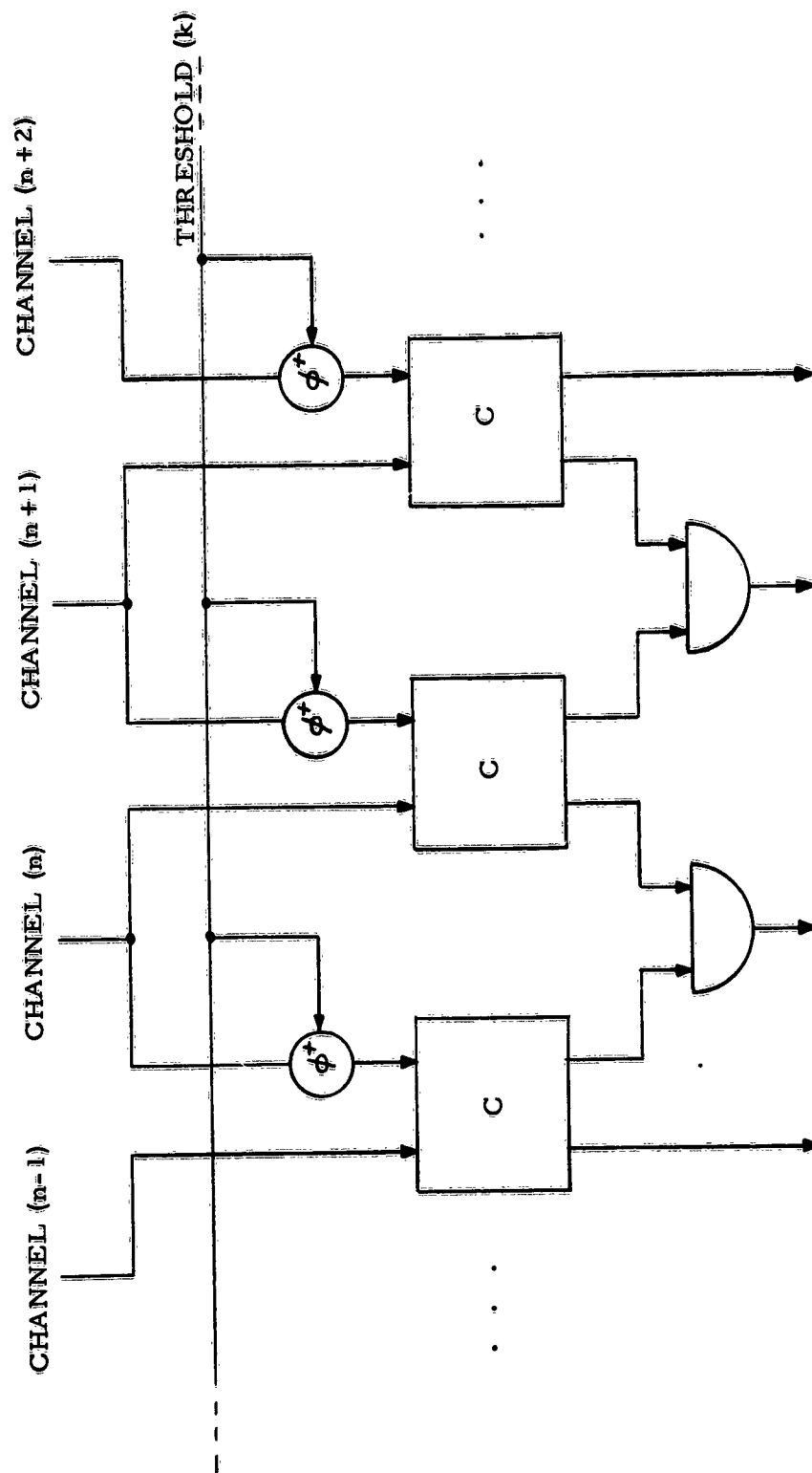


Figure 7. Spectrum Peak Picker

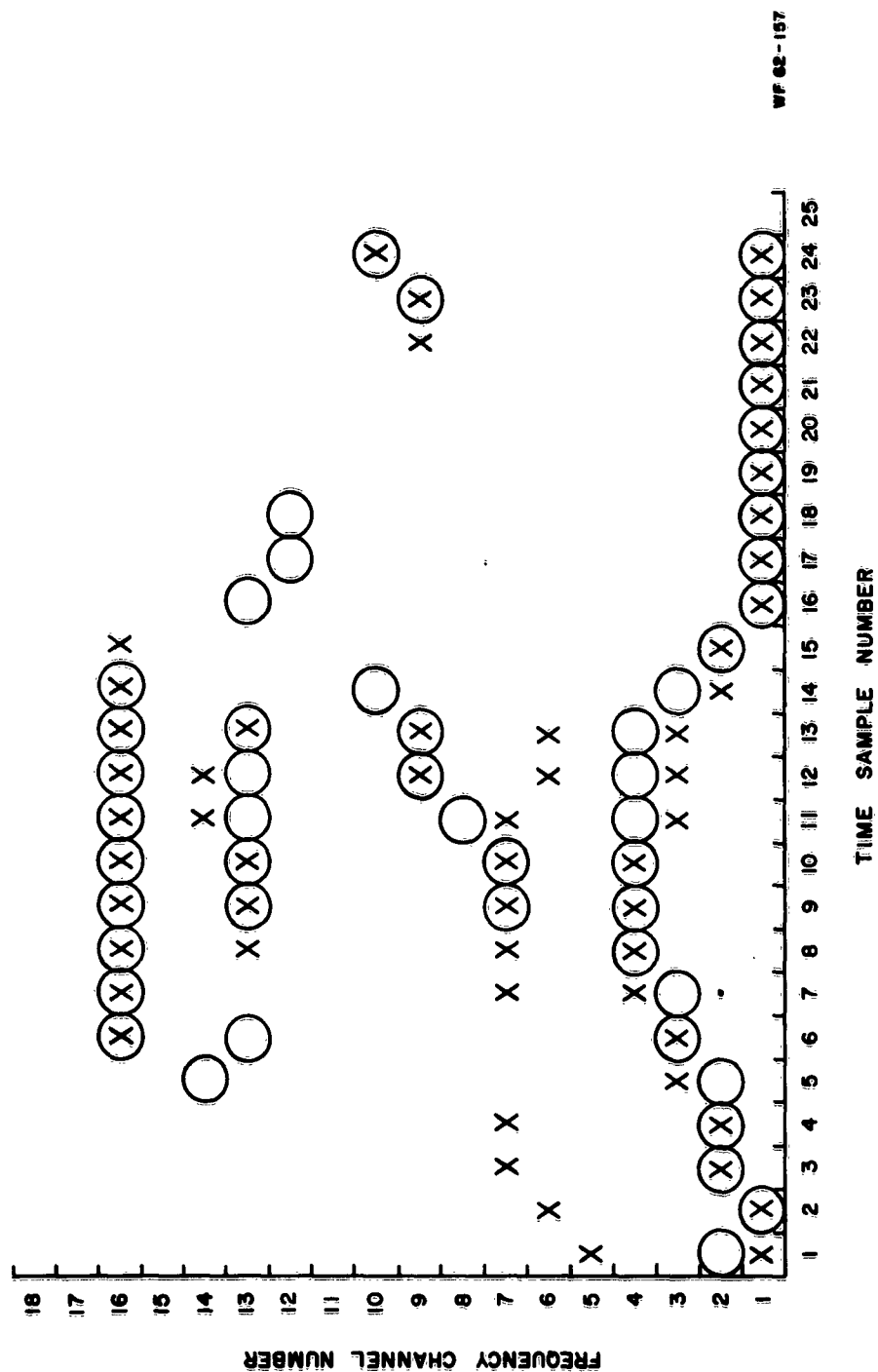


FIGURE 8. Formant Indications Provided by a Spectrogram (X) and a Peak Picker (O) for the Word "ONE"

$$M_{\nu} = \sum_{k=1}^{18} k^{\nu} f_k, \quad \nu = 0, 1, 2.$$

Approximations to these quantities have been obtained through the use of resistive adders. Actually, because of the relatively low skirt selectivity of the filters in the spectrum analyzer, the weightings employed in the equipment have been changed slightly (from k^{ν}) to compensate for the overlap between channels. The adjustment was made to produce an appropriate output of each moment extractor for a sinewave input. The resulting M_0 extractor produces an essentially constant output as a constant amplitude sinewave input is tuned over the 18 channels, and the M_1 and M_2 extractors produce outputs as indicated in Figure 9.

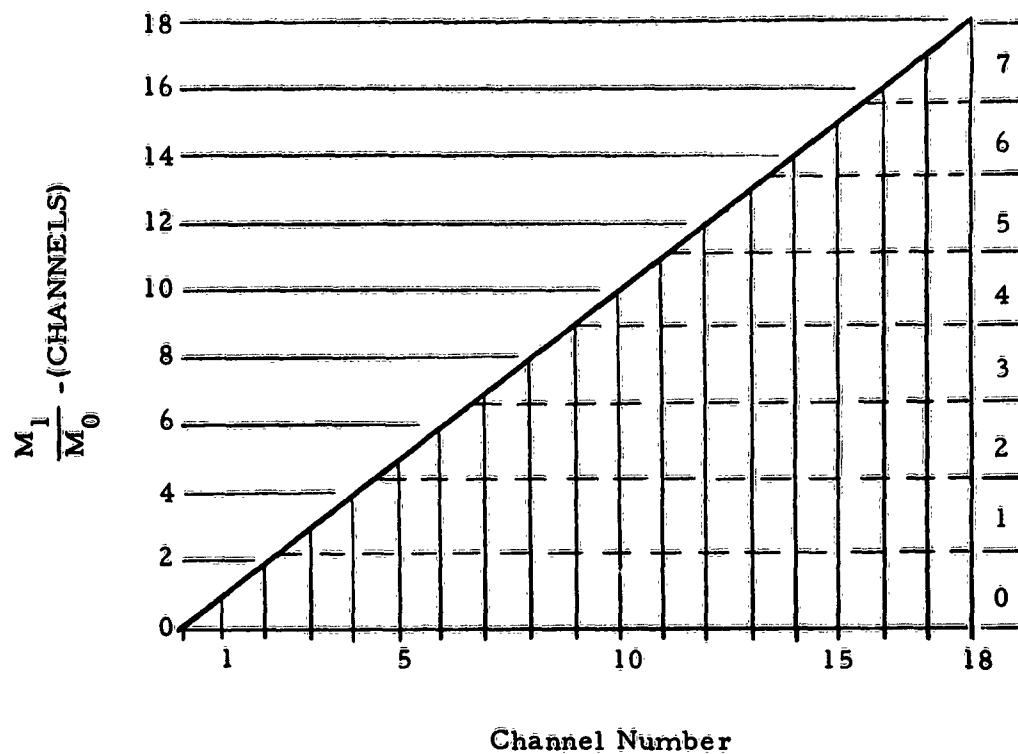
The segment boundary indicator, ΔS , is designed to detect changes in speech signals which correspond to transitions between speech sounds. The current method under study consists of adding the magnitudes of the derivatives of envelopes of the filter bank outputs:

$$\Delta S = \sum_{k=1}^{18} \left| \frac{df_k}{dt} \right|$$

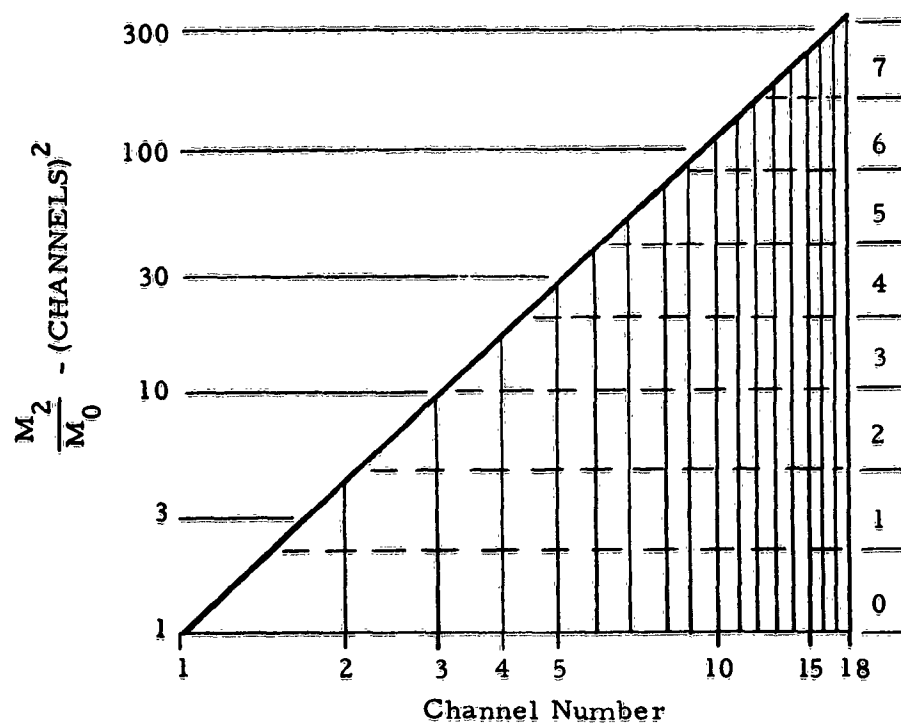
Although comprehensive testing of transcription methods employing this means of speech segmentation have not been completed, the boundaries created by thresholding ΔS do provide a reasonable correspondence between speech segments and utterances of phonetic language elements. This segmentation is illustrated in Table 4 for the words "Two Three". The speech segments indicated in this Table have been obtained by quantizing ΔS into 8 levels, and regarding the occurrence of the second or higher levels as transitions.

Two more parameters, a voicing indication (V) and pitch (F) are available in the current experimental equipment, but were not used in this study since most of the data obtained was for voiced sounds, and pitch provides little additional speech information.

(a) M_1



(b) M_2



3 BIT QUANTIZATION

Figure 9. Outputs of First and Second Spectrum Moment Extractors (M_1 M_2) for Sinewave Inputs At Different Frequencies.

TABLE 4. SEGMENTATION OF "TWO THREE" WITH THE PARAMETER ΔS

V	Peak Pattern (P)																		ΔS	Segment Identification
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18		
0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	111	T
0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	111	
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	010	
1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	111	...
1	0	0	1	0	0	0	0	1	0	1	0	0	1	0	0	1	0	0	100	
1	0	0	1	0	0	0	0	1	0	1	0	0	1	0	0	1	0	0	011	
1	0	0	1	0	0	0	0	1	0	1	0	0	1	0	0	1	0	0	010	OO
1	0	1	0	0	0	0	0	1	0	1	0	0	1	0	0	1	0	0	000	
1	0	1	0	0	0	0	0	1	0	1	0	0	1	0	0	1	0	0	000	
1	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	000	
1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	000	
1	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	
1	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	
1	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	
1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	000	
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	000	
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	001	
1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	001	
SILENCE																				
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	000	TH
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	000	
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	001	
1	0	1	0	0	0	1	0	1	0	0	0	0	1	0	0	1	0	0	111	...
1	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	011	
1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	000	
1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	001	UR
1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	001	
1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	001	
1	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	010	EE
1	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	000	
1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	000	
1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	001	
1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	000	
1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	000	
1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	000	
1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	000	
1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	000	
1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	000	
1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	000	
1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	001	
1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	000	
1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	001	

Note: Transition intervals are indicated by "...".

As indicated in Figure 5 (and discussed in Section 3.1.2 below) each extracted parameter is quantized to permit its representation as a binary number in a data format conversion unit. A digital commutation converts the binary numbers resulting for all parameters into a sequence of 5-bit samples suitable for direct insertion into the Recomp II computer. The pattern of parameter values resulting from a single speech sample is fed into one computer word location (capacity 40 bits). For convenience, the sampling interval, Δ , has been chosen equal to the time it takes to complete one complete drum revolution in the computer, thus producing 60 speech samples per second. Since the computer has a capacity of 4000 words, up to approximately a minute of continuous speech can be processed at one time.

The signals produced at any point in the block diagram in Figure 5 are available for display. For instance, the (sampled) output of the peak picking unit can be displayed on an oscilloscope as an intensity modulated sawtooth waveform with 60 sweeps per second. This display can be recorded to produce a representation of a segment of speech as a sequence of "instantaneous" spectra. As illustrated in Figure 10, an easily interpreted segment of one second of speech is economically, permanently and conveniently stored in a single 3" by 4" print. In this recording, a spectrum peak in a given channel is indicated by a white mark in that channel. Three channels are represented between each adjacent pair of horizontal grid lines in the photograph. Another parameter, the voicing indication, has been recorded in this photograph in the two positions above the top grid line in the photograph. Until the data format conversion equipment became available toward the end of this project, photographs of this type were used to obtain data for the parameter space formed by spectral peaks and the voicing indication. Before the peak picker itself was constructed, the program outlined in Appendix I was used to simulate its operation on the Recomp II computer. Quantized sequences of sample spectra were obtained as inputs to this program through the courtesy of Mr. C. P. Smith of the AFCRL Communications Laboratory.

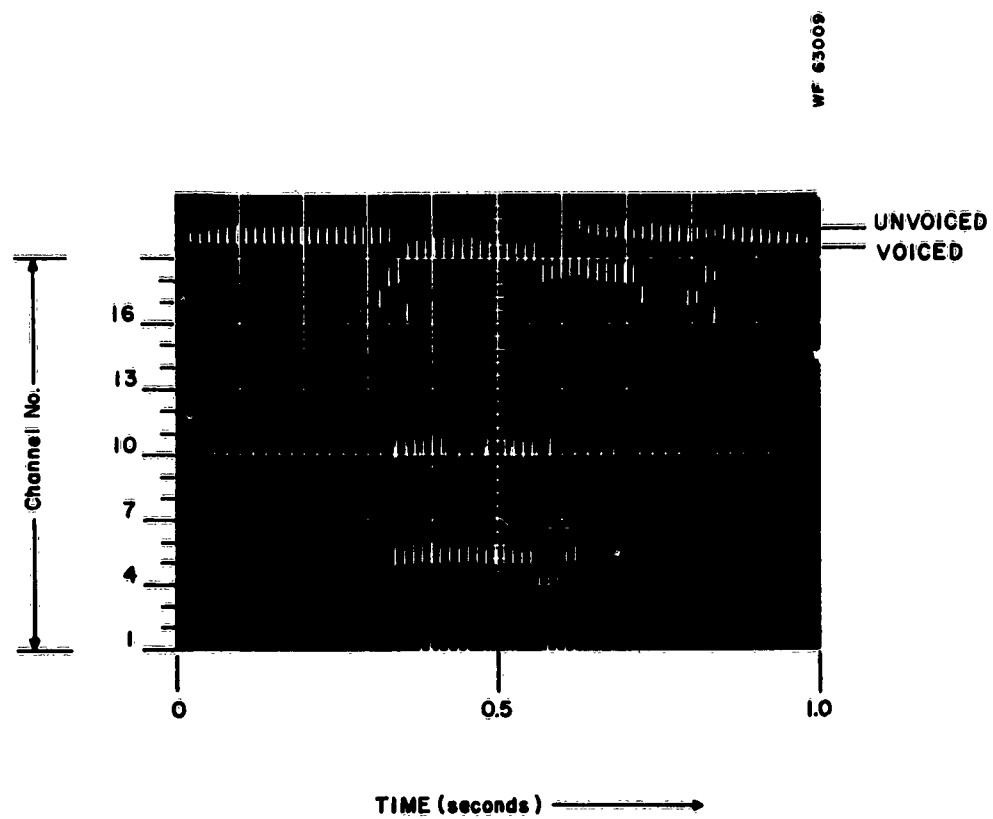


FIGURE 10. Spectrum Peak Picker Output Representation Of the Spoken Word "ASK"

3.1.2 Parameter Quantization

As remarked in Section 2.2, quantization of parameter values constitutes an integral part of the way in which a parameter is defined. Unfortunately, however, time and equipment limitations have prevented a thorough study of the effects of several different quantizations which would be reasonable for the parameters studied. Specifically, the way in which each parameter has been quantized for experiments reported for this study is indicated in Table 5. The extractor for each of the parameters E , E_0 , M_1 , and M_2 was designed to produce outputs between 0 and 6 volts, and the eight quantizing levels were set at 0.75 volt intervals, so that a linear quantization of these five parameters was obtained. While this representation is quite reasonable for E , E_0 , M_1 , and M_2 , it is not satisfactory for M_2 . Significantly better resolution would be obtained for this latter parameter with logarithmic spacing of quantizing levels (as indicated in Figure 9). However, since only one A/D converter was available in time for use on this project, a compromise was made in favor of the linear spacing. The effects of different quantization of the spectrum moments are discussed in Section 5 in the light of experimental results reported in the remainder of this section.

TABLE 5. SPEECH PARAMETER QUANTIZATION

<u>Parameter</u>		<u>Maximum Number of Different Parameter Values</u>	
<u>Symbol</u>	<u>Description</u>	<u>Number</u>	<u>Bits</u>
	Spectral peak pattern	$\approx 2^{13}$	13
V	Voicing indication	2	1
E	Input speech amplitude	8	3
E_0	Normalized speech amplitude	8	3
M_0	Spectrum area	8	3
M_1	Spectrum First Moment	8	3
M_2	Spectrum Second Moment	8	3

3.2 DISTRIBUTION OF SPEECH SOUNDS IN PARAMETER SPACE

To ascertain the distribution of speech sounds in the parameter spaces created by combinations of the parameters listed in Table 3, data has been collected from three speakers using the equipment described in the preceding section. Results are reported here for the eleven vowel sounds listed in Table 1. To obtain a representative set of speech samples of these sounds, the following procedure has been followed for each speaker.

The speaker was asked to read a word list consisting of eleven words. Each word on the list was chosen so that one of the eleven vowel sounds would be spoken during the utterance of the word, if the word were pronounced "properly". To obtain a reasonably large number of independent samples of sounds within a varying environment, each speaker was presented with 3 different word lists, at 5 different times, within an interval of several days. The three word lists are contained in Table 6. This procedure produced a magnetic tape recording of 15 utterances of each of the 11 vowel sounds by each of the three speakers -- 495 utterances in all.

The next step consisted of playing back the recorded word lists into the speech processing equipment. This resulted in a sequence of sample patterns of parameter values representing the speech, stored in the Recomp computer. This sequence of patterns was then typed out, and the intervals corresponding to the vowel sounds identified. A human observer made the identification while listening to the original speech recording, and using the following rough guidelines:

- a) Use only those patterns which occur within intervals of speech comprising readily identifiable sounds.
- b) Use primarily those patterns which either persist over several samples or which change slowly over an interval of unchanging sound.

The typed representation of this sequence of patterns and a typical assignment of patterns to a sound are illustrated in Table 7, using the word "Neck", and the interval associated with the sound EH (e). Interpretations of the binary representation of the parameters are given in Table 8. All parameters have been represented in binary form for convenience only -- octal and decimal representations are also available with the Recomp computer.

TABLE 6 - THREE WORD LISTS EMPLOYED FOR VOWEL SOUNDS

<u>Word List Number One</u>	<u>Word List Number Two</u>	<u>Word List Number Three</u>
1. HID	1. BASK	1. TOE
2. COOK	2. NOT	2. VAST
3. NECK	3. FRAUD	3. SET
4. BOG	4. NOOK	4. POD
5. FALL	5. EARL	5. TRUE
6. WOO	6. MOOSE	6. AWE
7. OAK	7. RUB	7. DEED
8. TURF	8. NO	8. FUSS
9. AS	9. DEATH	9. GOOD
10. BEAN	10. HEAP	10. PERK
11. BUD	11. SIT	11. RIB

Speech Parameter Pattern													
Time Sample No.	V	P						M ₀	M ₁	M ₂	E	E ₀	
1	1	010	000	000	000	100	000	101	010	001	001	011	
2	1	010	000	000	000	100	000	110	001	000	011	110	
3	1	010	000	000	100	100	000	101	010	001	011	100	
4	1	010	000	000	100	010	000	111	100	010	011	100	
5	1	001	000	000	010	010	000	111	101	011	011	100	
6	1	100	100	000	010	010	000	111	101	011	011	100	
7	1	000	100	000	100	010	100	111	100	011	100	100	
8	1	000	100	000	100	100	100	111	101	011	101	100	
9	1	000	100	000	000	100	100	111	101	011	100	011	
10	1	001	000	000	001	000	000	100	010	010	001	001	EH(E)
11	1	001	000	000	000	000	000	001	000	000	000	001	
12	1	001	000	000	000	100	000	010	001	000	000	001	
13	1	001	000	000	000	000	000	001	000	000	000	000	
14	1	001	000	000	000	000	000	001	000	000	000	000	
15	1	001	000	000	000	000	000	001	000	000	000	000	
16	0												
17	0												
18	0												
19	0												
20	0												
21	0												
22	0	001	000	000	000	000	000	000	000	000	000	011	
23	0	001	000	000	001	000	010	110	101	011	000	001	
24	0	001	000	000	000	100	001	010	010	010	000	001	
25	0	000	000	000	000	100	000	001	000	001	000	000	

Table 7. Representation of the Word "NECK" in Parameter Space

Parameter	Bit Position(s)	Value	Interpretation
Voicing (V)	1	0 1	Unvoiced Voiced
Spectral Peaks (p)	2-19	1 0	Peak Present in a Frequency Channel Peak not present in a Frequency Channel
Spectrum Area (M_0)	20-22	3-digit Binary Number	Magnitude of Corresponding Parameter
Spectrum Mean (M_1)	23-25		
Spectrum Second Moment (M_2)	26-28		
Input Speech Amplitude (E)	29-31		
Normalized Speech Amplitude (E_0)	32-34		

Table 8. Binary Representations of Speech Parameters

The remaining step in the processing of speech data to obtain an estimate of the distribution of sounds in parameter space, consists of listing each different pattern selected by the above procedure, along with the number of times the pattern occurred within each sound. From such a histogram the regions in parameter space corresponding to each sound, and the overlap between these regions can be ascertained.

3.2.1 Parameter Space Usage

The number of sample patterns processed to obtain a picture of the distribution of vowel sounds, and the number of different patterns which arose from each sound individually, and from all vowel sounds, are given in Table 9. As indicated in this table, most of the data processing has been performed for the two parameter spaces formed by (1) considering the location of spectral peaks alone, and (2) considering spectral peaks and the first three spectrum moments: M_0 , M_1 and M_2 . These two spaces will be called "peak space" and "peak-moment space", respectively. Since 60 speech samples are extracted each second, we see from this table that a total of approximately 28, 23, and 27 seconds of spoken vowel sounds were processed to obtain the data for the three speakers. Thus, an average of approximately 2.5 seconds of each sound was obtained for each speaker. Since each sound was uttered 15 times by each speaker, an average of approximately 10 sample patterns were obtained from each utterance of a vowel sound.

Perhaps the first question which arises in considering such a collection of data is whether enough samples have been obtained to warrant acceptance of the distribution of sounds in parameter space provided by the samples, as an accurate estimate of the distribution which would be observed if an unrestricted number of speech samples were processed. In an attempt to obtain at least a partial answer to this question it has been conjectured* that the number of different speech parameter patterns, N_p , which occur within an interval of speech, T , tends to vary according to a parametric space usage curve:

$$N_p = N_0 [1 - e^{-\alpha \Delta}]$$

*[9].

Sound	Speaker No. One			Speaker No. Two			Speaker No. Three		
	No. of Patterns		No. of Samples	No. of Patterns		No. of Samples	No. of Patterns		No. of Samples
	Peak- Space	Moment Space		Peak- Space	Moment Space		Peak- Space	Moment Space	
A(ae)	211	35	85	157	49	98	181	45	113
AH(a)	180	68	105	148	68	107	136	43	82
AW()	214	37	72	177	64	109	178	41	72
EE(i)	172	35	91	147	32	88	189	22	83
EH(e)	103	48	79	88	38	75	114	50	98
I(I)	106	43	75	93	51	74	113	29	80
O(O)	136	60	101	95	47	95	120	27	57
OO(u)	168	31	57	161	15	27	160	16	52
U(U)	102	41	74	99	46	75	141	26	78
UH()	121	48	88	91	43	72	98	27	56
UR()	164	39	81	129	31	74	203	17	68
All Vowel Sounds	1677	308	787	1385	339	793	1633	186	657

TABLE 9 - NUMBER OF SAMPLES AND PATTERNS OBTAINED FROM
THREE SPEAKERS AND VOWEL SOUNDS

where N_0 is some number which is less than or equal to the total number of patterns which can possibly occur, α is a constant reflecting the rate at which the number of different patterns encountered grows as the length of speech observation interval is increased, Δ is the sampling interval, and n is the number of samples obtained in the interval, T . The quantities N_0 and α are determined by the parameter space and the speech source.

A thorough exploration of the parameter space usage curve for the spaces formed by the speech parameters considered during this study has not been possible. However, based on a few points obtained for a single speaker, the two values $N_0 = 400$ and $\alpha = .05$ provide a reasonably good fit for the generation of spectral patterns alone, i. e., for the parameter space formed by positions of spectral peaks. Since the total number of spectral peak patterns which can ever occur has been found to be approximately 2^{13} , it appears that not much more than five percent of the points in this parameter space would ever be used by vowel sounds, no matter how long an interval of speech is considered. A further indication provided by $N_0 = 400$ and $\alpha = .05$ is that at least 75 percent of all sample patterns produced by vowel sounds would be matched by the 308 peak patterns generated by the 1677 samples taken from speaker number one (Table 9).

Although great reliance should not be placed on the parameter space usage curve until further study is performed, it is encouraging to note that 81 percent of the test samples processed for speaker number one in the transcription experiments (described in Section 4) matched one of the 308 patterns generated by the vowel sound data - a discrepancy of six percent between observed and predicted relative frequency of matching.

In the interest of avoiding undue bulk, the complete histogram for each sound has not been included in this report. Rather, a few salient characteristics of the sound distributions are described. However, as an indication of the type of distribution obtained, the complete distribution of a single sound for a single speaker has been included in Table 10 for peak space, i. e., the parameter space formed by spectral peaks alone.

TABLE 10. COMPLETE HISTOGRAM FOR THE SOUND EE(i) IN PEAK SPACE (SPEAKER NUMBER ONE)

Rank Order	Pattern (P)																		No. of Occurrences
	Channel Number																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	28
2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	16
3	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	16
4	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	15
5	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	12
6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	11
7	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	8
8	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	7
9	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	7
10	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	6
11	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	6
12	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	4
13	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	4
14	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	3
15	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	3
16	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	3
17	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	2
18	1	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	2
19	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	2
20	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	2
21	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
22	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
23	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
24	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1
25	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1
26	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1
27	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1
28	1	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	1
29	0	1	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	1
30	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1
31	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	1
32	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	1
33	1	0	0	0	0	1	0	0	0	0	0	1	0	1	0	1	0	0	1
34	1	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	1	1
35	0	1	0	0	0	0	0	0	0	1	0	1	0	1	0	1	0	0	1

The high frequency of occurrence of some patterns within a sound suggests that a large percentage of all speech samples might correspond to a very small number of patterns of parameter values. As an indication of both the region in parameter space occupied by the vowel sounds, and the degree to which speech may be represented by a small number of patterns, the ten most frequently occurring patterns have been listed in Appendix II for each of the vowel sounds, in peak space and peak-moment space for speaker number one. The coverage of all samples obtained from the vowel data provided by these 110 patterns, is summarized in Table 11 for each of the three speakers.

As a final indication of parameter space usage, the distribution of speech sample spectra with respect to the number of spectral peaks is shown in Figure 11. The graphs in this figure indicate that the majority of speech samples (of vowel sounds) produce spectra with three or four peaks. Further, the relative frequency of occurrence of a given number of peaks appears to be approximately the same for the three speakers.

In the course of this study, some thought has been directed to the question of whether the parameter space usage can be reduced (without increasing overlap between speech sounds) by some sort of "warping" performed after patterns of parameter values are obtained. One technique for attempting to accomplish this reduction has been investigated for peak space. Specifically, it has been conjectured that the most important characteristic of formant positions consists of the ratio of the second and higher formant frequencies to the first formant frequency. If this is the case, and if the vocoder filters are logarithmically spaced, then allowable variations in formant positions representing a given speech sound would consist of "rigid" shifts of the peak-picked spectra. To test this idea, a program has been written which maps peak-picked spectra into a subset as determined in the following way. The first sample occurring in a speech recording is installed as the initial member of an "intermediate reference library". Each successive new spectrum arising in the speech recording is compared with each of the spectra in this library. If a new spectrum meets a criterion of "closeness" to any one of the members of the library, then the new spectrum is associated with that reference library member. If the criterion is not met for any of the library members, then the new spectrum is installed as a new member of the library. A criterion of closeness based on the notion that slight, rigid shifts in formant positions are allowable, has been tested using the Recomp II computer. A description of the program is given in Appendix III. These tests indicate that while the number of different library members tends to level off at a few hundred as speech samples are processed, more work must be done to relate the criterion of closeness to the speech sounds themselves.

Fraction of Vowel Samples Covered by 10 Most Frequently Occurring Patterns						
Speech Sound	Speaker No. One		Speaker No. Two		Speaker No. Three	
	Peak-Moment Space	Peak Space	Peak-Moment Space	Peak Space	Peak-Moment Space	Peak Space
A	.35	.75	.30	.69	.26	.60
AH	.35	.56	.24	.40	.34	.52
AW	.42	.67	.28	.45	.48	.64
EE	.36	.74	.27	.74	.37	.91
EH	.29	.56	.24	.64	.20	.49
I	.33	.61	.23	.46	.22	.57
O	.25	.49	.25	.42	.46	.74
OO	.63	.84	.83	.98	.60	.97
U	.27	.56	.22	.56	.38	.69
UH	.28	.58	.29	.54	.46	.79
UR	.48	.76	.40	.79	.55	.96
Average of Vowels	.37	.65	.32	.61	.39	.72

Table 11. Fraction of Vowel Samples Covered by the Ten Most Frequently Occurring Patterns

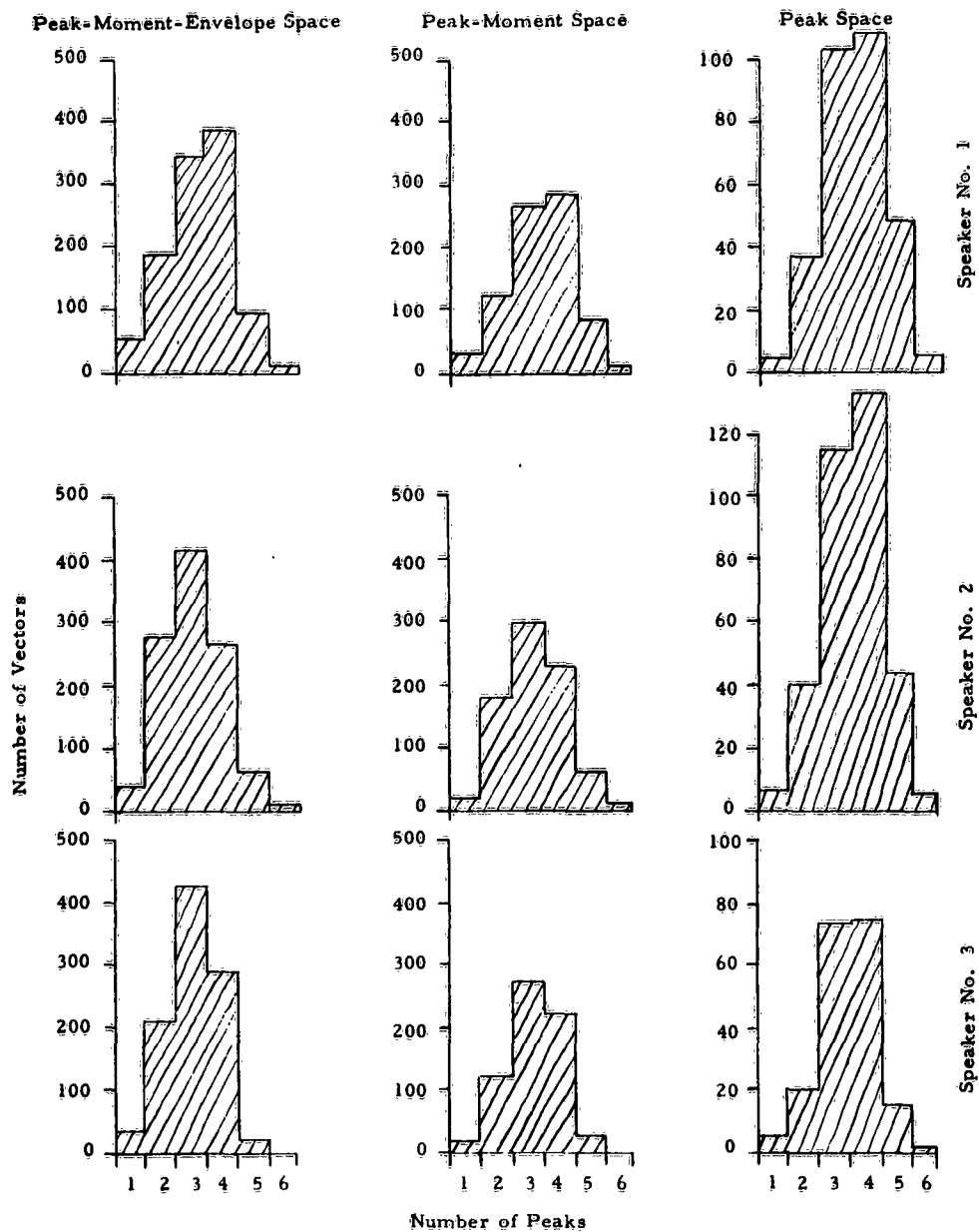


Figure 11. Reference Library Structure (By Number of Spectrum Peaks) for Three Speakers and Three Combinations of Parameters (Vowel Sounds Only)

3.2.2 Overlap Between Speech Sounds in Parameter Space

As discussed in Section 2.3, it is desired that a parameter space be constructed in such a way that any given pattern of parameter values always arises from the same sound. If this goal is achieved, then parameter space may be partitioned into non-overlapping regions with each region corresponding to exactly one speech sound. However, as is the case with many other pattern recognition problems exhibiting wide variations in manifestation of the classes involved, complete absence of overlap between speech sounds (or any other language-elements) will very likely never be attained with any parameter space.

The degree to which a given set of parameters can be expected to provide adequate separation of speech sounds can be estimated in several ways. One of the more informative ways would be to calculate the probability that any given speech sound will be designated as one of the other speech sounds, when parameter space is partitioned in a way which tends to minimize this quantity. As discussed in Section 2.3, many methods exist by which such a partitioning can be approximated. If estimates of the probabilities $\{P_i(\chi | S_i)\}$ are available then the maximum likelihood method of partitioning can be applied with these estimates. We have applied this method of partitioning parameter space, using the histograms for each sound as estimates of $\{P_i(\chi | S_i)\}$. Letting χ_k denote the k-th distinct pattern of parameter values produced by samples of all speech sounds (limited to vowels for the data reported here), the probability that a single sample of speech corresponding to the i-th speech sound will be associated with the j-th speech sound, α_{ij} , can be written (for $i \neq j$)

$$\alpha_{ij} = \frac{1}{n_i} \sum_{k=1}^n \beta_{ki} \gamma_{ij}^{(k)}$$

where β_{ki} = the number of occurrences of the pattern χ_k within intervals of speech corresponding to the i-th speech sound,

$$\gamma_{ij}^{(k)} = \begin{cases} 1 & \text{if } \frac{\beta_{ki}}{n_i} < \frac{\beta_{kj}}{n_j} \\ 0 & \text{otherwise} \end{cases}$$

n_i = the number of samples arising from the i -th speech sound

and

$n = \sum_{i=1}^N n_i$ = the total number of available speech samples.

Using the data collected for the eleven vowel sounds, the $\{\alpha_{ij}\}$ matrices are given in Tables 12, 13, and 14 for three speakers, and for peak space and peak-moment space. The diagonal elements in these matrices, α_{ii} , indicate the estimated probability of correct classification. In some cases, the most likely confusions occur between speech sounds which are often interchanged by speakers. Many people, for instance, use A instead of EH for the first vowel sound in the word "HELLO". The result of interchanging these sounds would be a speech transcription with a distorted accent. For non vowel sounds, of course, interchanges can produce more detrimental results.

As with parameter space usage, the adequacy of the number of available samples may be questioned when considering such overlap matrices. Although such devices as the parameter space usage curve introduced earlier may be employed in an attempt to answer this question, time has not allowed for this type of study during this project.

The pairwise overlaps indicated by the $\{\alpha_{ij}\}$ matrices, do not show the extent to which more than two speech sounds ever give rise to the same pattern of parameter values. The number of different patterns with p spectral peaks which ever occur within intervals of speech corresponding to k sounds is indicated in Table 15 for each of three speakers and for peak space. The entries in this table, coupled with the $\{\alpha_{ij}\}$ matrices suggest that although approximately 30 percent of the different patterns in peak space occur, at different times, within intervals of speech corresponding to different sounds, most occurrences of these patterns are associated with a single sound.

As a final indication of the degree to which the vowel sounds produce different patterns in peak space, we have constructed bar graphs, called spectral profiles, which show the percentage of sample patterns

Sound Spoken	Probability (Percent) of Being Recognized As											*
	A	AH	AW	EE	EH	I	O	OO	U	UH	UR	
A	80	02	02	00	08	03	00	00	01	04	01	EH
AH	02	92	03	00	01	00	00	00	00	02	00	AW
AW	08	04	79	00	05	01	01	00	01	01	01	A
EE	00	00	00	98	00	01	00	02	00	00	00	OO
EH	15	02	02	00	58	17	01	00	03	00	02	I
I	03	00	00	00	05	74	02	00	11	02	04	U
O	00	02	04	00	01	02	47	00	10	08	25	UR
OO	00	00	00	04	01	00	04	88	04	00	01	EE
U	00	00	00	00	00	06	07	00	71	10	07	UR
UH	06	03	02	00	10	02	03	00	04	67	02	A
UR	01	00	00	00	01	03	02	01	04	02	88	U

Peak Space

Sound Spoken	Probability (Percent) of Being Recognized As											*
	A	AH	AW	EE	EH	I	O	OO	U	UH	UR	
A	88	01	01	00	06	01	00	00	01	03	00	EH
AH	01	94	04	00	00	00	01	00	00	01	00	AW
AW	02	01	91	00	03	00	00	00	00	01	00	EH
EE	00	00	00	99	00	00	00	00	00	00	00	
EH	09	00	00	00	82	07	00	00	02	01	00	A
I	00	00	00	00	04	89	00	00	05	02	01	U
O	01	01	01	00	01	01	81	00	04	02	07	UR
OO	00	00	00	01	00	00	03	97	01	00	00	O
U	00	00	00	00	00	03	02	00	89	02	04	UR
UH	05	01	02	01	05	02	00	00	02	84	00	A
UR	01	00	00	00	01	00	04	00	02	01	92	O

Peak-Moment Space

Table 12. Estimated Relative Frequency of Correct and Misclassification of Vowel Sounds for Speaker Number One, and Two Parameter Spaces.

* Most Likely Confusion

Sound Spoken	Probability (Percent) of Being Recognized As											*
	A	AH	AW	EE	EH	I	O	OO	U	UH	UR	
A	75	01	01	00	22	01	00	00	01	01	00	EH
AH	07	68	12	00	01	00	03	01	01	07	00	AW
AW	06	03	74	03	02	01	07	01	01	03	01	O
EE	00	00	00	99	00	00	00	00	00	00	00	
EH	02	00	08	00	81	07	00	00	00	02	00	AW
I	00	00	00	00	10	85	01	00	02	00	02	EH
O	00	00	10	00	02	02	58	00	11	08	10	U
OO	00	00	00	00	00	00	02	96	02	00	00	U
U	00	00	01	01	01	01	10	15	62	02	07	OO
UH	00	08	04	00	06	01	00	00	01	79	01	AH
UR	00	00	00	00	01	06	09	02	04	02	77	O

Peak Space

Sound Spoken	Probability (Percent) of Being Recognized As											*
	A	AH	AW	EE	EH	I	O	OO	U	UH	UR	
A	91	01	02	00	06	00	00	00	00	00	00	EH
AH	02	89	04	00	00	00	01	00	00	04	00	AW
AW	01	08	80	00	01	00	07	00	00	02	01	AH
EE	00	00	00	99	00	00	00	00	00	00	00	
EH	06	01	03	00	88	00	00	00	00	02	00	A
I	01	00	00	00	01	97	00	00	00	01	00	EH
O	00	00	03	00	00	01	86	01	03	02	04	UR
OO	00	00	00	00	00	00	02	96	02	00	00	U
U	00	00	00	00	00	01	09	05	82	00	03	O
UH	00	00	03	00	00	00	00	00	00	96	01	AW
UR	00	00	00	00	00	03	02	01	02	00	92	I

Peak-Moment Space

Table 13. Estimated Relative Frequency of Correct and Misclassification of Vowel Sounds for Speaker Number Two, and Two Parameter Spaces.

* Most Likely Confusion

Sound Spoken	Probability (Percent) of Being Recognized As											*
	A	AH	AW	EE	EH	I	O	OO	U	UH	UR	
A	89	04	00	00	04	00	00	00	00	03	00	EH
AH	03	49	18	00	00	01	03	00	04	07	15	AW
AW	04	10	62	00	00	02	10	00	03	09	00	AH
EE	00	00	00	99	00	00	00	00	00	00	00	
EH	18	02	03	00	52	19	04	01	00	00	04	I
I	00	01	00	04	09	70	02	00	01	00	13	UR
O	00	07	12	00	00	08	64	00	02	00	08	AW
OO	00	02	02	08	00	02	01	83	01	00	00	EE
U	00	03	02	00	02	14	10	00	41	00	28	UR
UH	09	14	05	00	01	00	00	00	01	69	00	AH
UR	01	02	00	00	01	13	01	00	01	01	81	I

Peak Space

Sound Spoken	Probability (Percent) of Being Recognized As											*
	A	AH	AW	EE	EH	I	O	OO	U	UH	UR	
A	92	01	02	00	04	00	00	00	00	02	00	EH
AH	00	65	12	00	01	00	04	00	00	04	14	UR
AW	02	07	73	00	00	01	10	00	00	07	00	O
EE	00	00	00	98	00	02	00	00	00	00	00	I
EH	06	02	01	00	77	07	00	00	02	02	02	I
I	00	00	00	00	03	84	03	00	04	00	07	UR
O	00	02	08	00	00	02	77	00	05	01	04	O
OO	00	01	00	00	00	00	02	98	00	00	00	O
U	01	03	03	00	01	04	07	00	57	00	25	UR
UH	06	09	01	00	01	00	00	00	00	82	00	AH
UR	00	01	00	00	01	05	01	00	05	00	88	U

Peak-Moment Space

Table 14. Estimated Relative Frequency of Correct and Misclassification of Vowel Sounds for Speaker Number Three, and Two Parameter Spaces.

* Most Likely Confusion

Number of Sounds (k)	Number of Peaks (p)							Number of Peaks (p)							Number of Peaks (p)								
	Number of Peaks (p)						Totals	Number of Peaks (p)						Totals	Number of Peaks (p)						Totals		
	1	2	3	4	5	6		1	2	3	4	5	6		1	2	3	4	5	6			
1	0	17	65	80	38	5	205	1	3	17	74	110	40	4	248	1	1	5	41	48	14	0	109
2	3	13	24	17	6	0	63	2	2	13	20	17	3	0	55	2	0	2	11	13	2	0	28
3	1	4	5	7	3	0	20	3	2	5	14	7	1	0	29	3	0	5	13	9	0	0	27
4	1	3	5	1	1	0	11	4	1	3	2	0	0	0	6	4	1	3	7	2	0	0	13
5	0	0	3	3	0	0	6	5	0	2	3	0	0	0	5	5	1	1	1	2	0	0	5
6	0	0	0	1	0	0	1	6	0	0	0	0	0	0	0	6	0	3	0	0	0	0	3
7	0	0	0	1	0	0	1	7	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0
Totals	5	37	103	110	48	5		Totals	8	40	113	134	44	4		Totals	3	19	73	74	16	0	

Speaker Number One

Speaker Number Two

Speaker Number Three

Speaker Number One

Speaker Number Two

Speaker Number Three

Table 15. Number of Different Patterns with p Peaks which Ever Arise from k Speech Sounds

observed for a given speech sound, which possessed a peak in the k -th channel, $k = 1, 2, \dots, 18$. These profiles for the vowel sounds and the three speakers are shown in Figures 12, 13, and 14. Although information concerning combinations of peaks is not contained in these graphs, in some cases the locations of formants can be inferred. In the sound EE, for instance, the sum of samples containing a peak in either channels 1 or 2 accounts for essentially all samples; this is true for all three speakers. This verifies the well-known fact that EE produces a first formant in the frequency range 200 - 400 cps.

One is easily tempted to draw conclusions from the spectral profiles other than simply the approximate locations of formants as indicated by the peak picker. In the sound EE, for instance, the relative frequency of occurrence of a peak in the first channel may be used as an estimate of the likelihood that an utterance of this sound will produce a peak in that channel. Also, interpolation using the channel weightings indicated in the spectral profiles might be expected to provide a more accurate estimate of the formant locations for a given speaker. Further, in some cases (for instance the second formant in the sound OO), the sum of all weightings in a short frequency interval spanning no more than two or three channels, and surrounded by channels with all-zero weightings, may provide an indication of formant strength.

From the data obtained on vowel sounds, it can be concluded that peak-space provides a means of representation which achieves adequate separation of all vowel sounds except those sounds which are perhaps the most difficult for a human to distinguish between. In peak-moment space, further, but not complete, separation is achieved. It is very likely that the reason for incomplete separation of vowel sounds in peak-moment space is due solely to the way in which the parameters M_1 and M_2 have been quantized. If these parameters are quantized properly (as indicated in Figure 9), then essentially complete separation of vowel sounds is expected, and further, the peak-moment parameter space usage could very easily be less than that reported in Section 3.2.1 above.

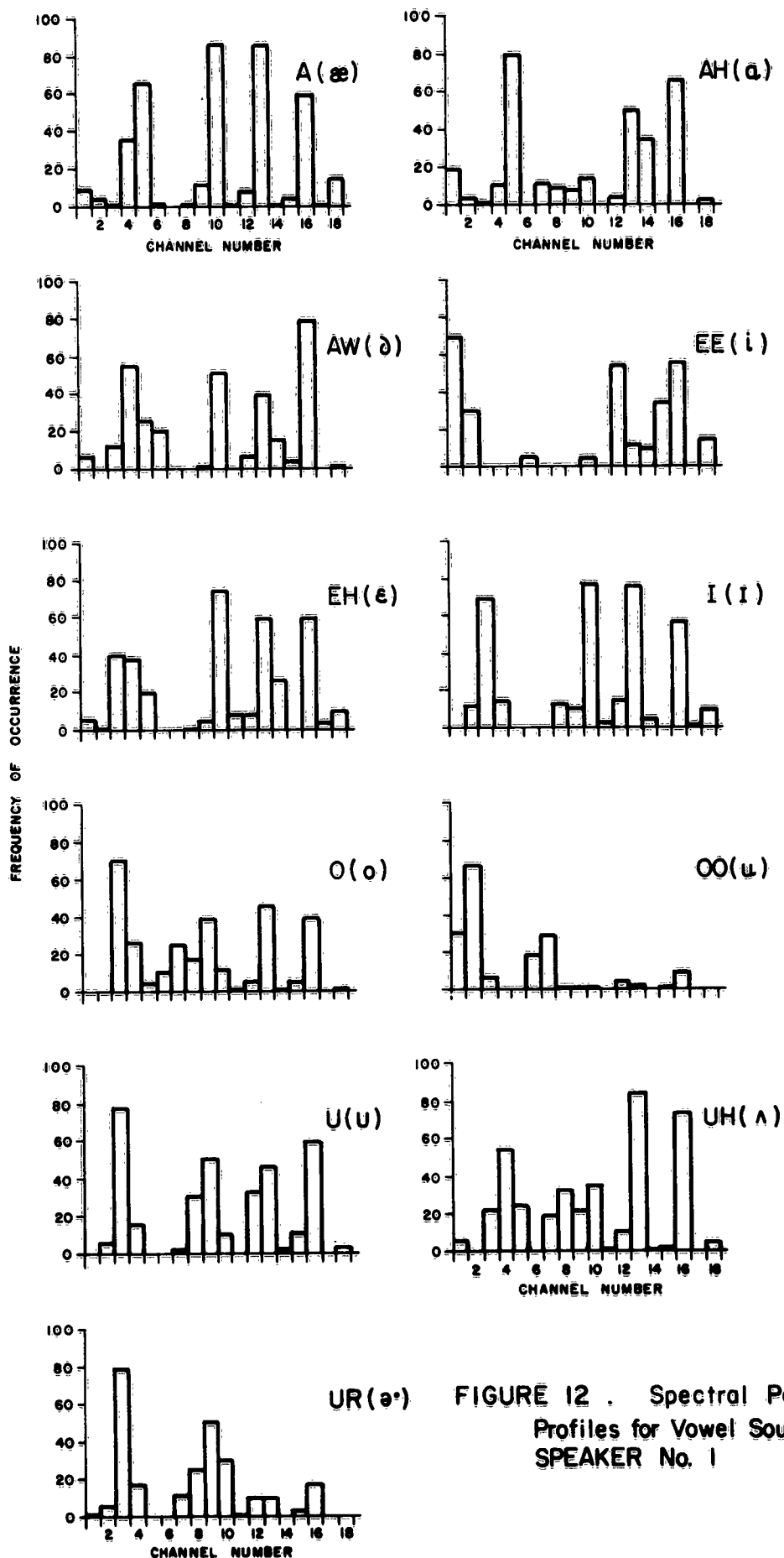
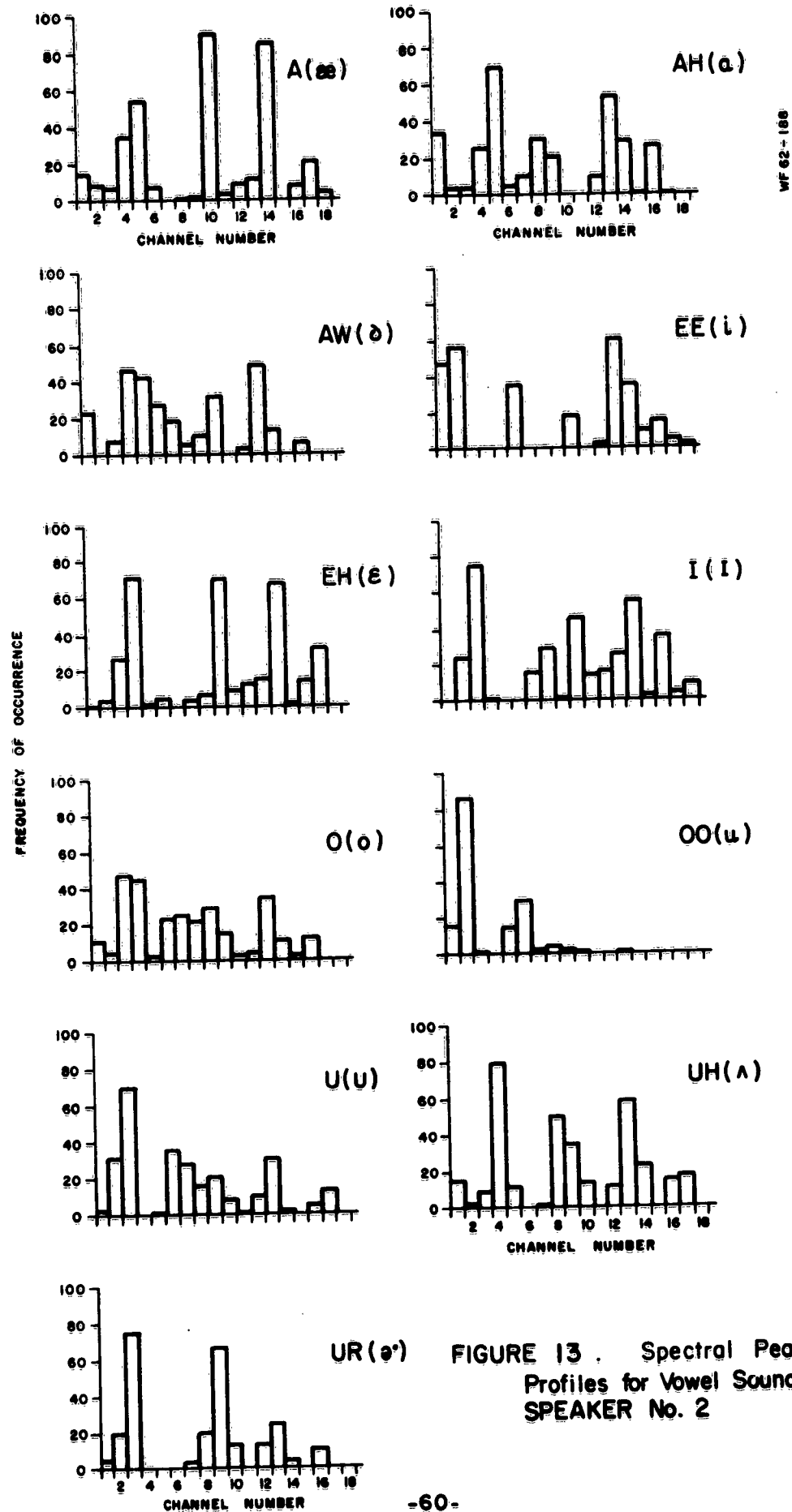


FIGURE 12 . Spectral Peak Profiles for Vowel Sounds, SPEAKER No. 1



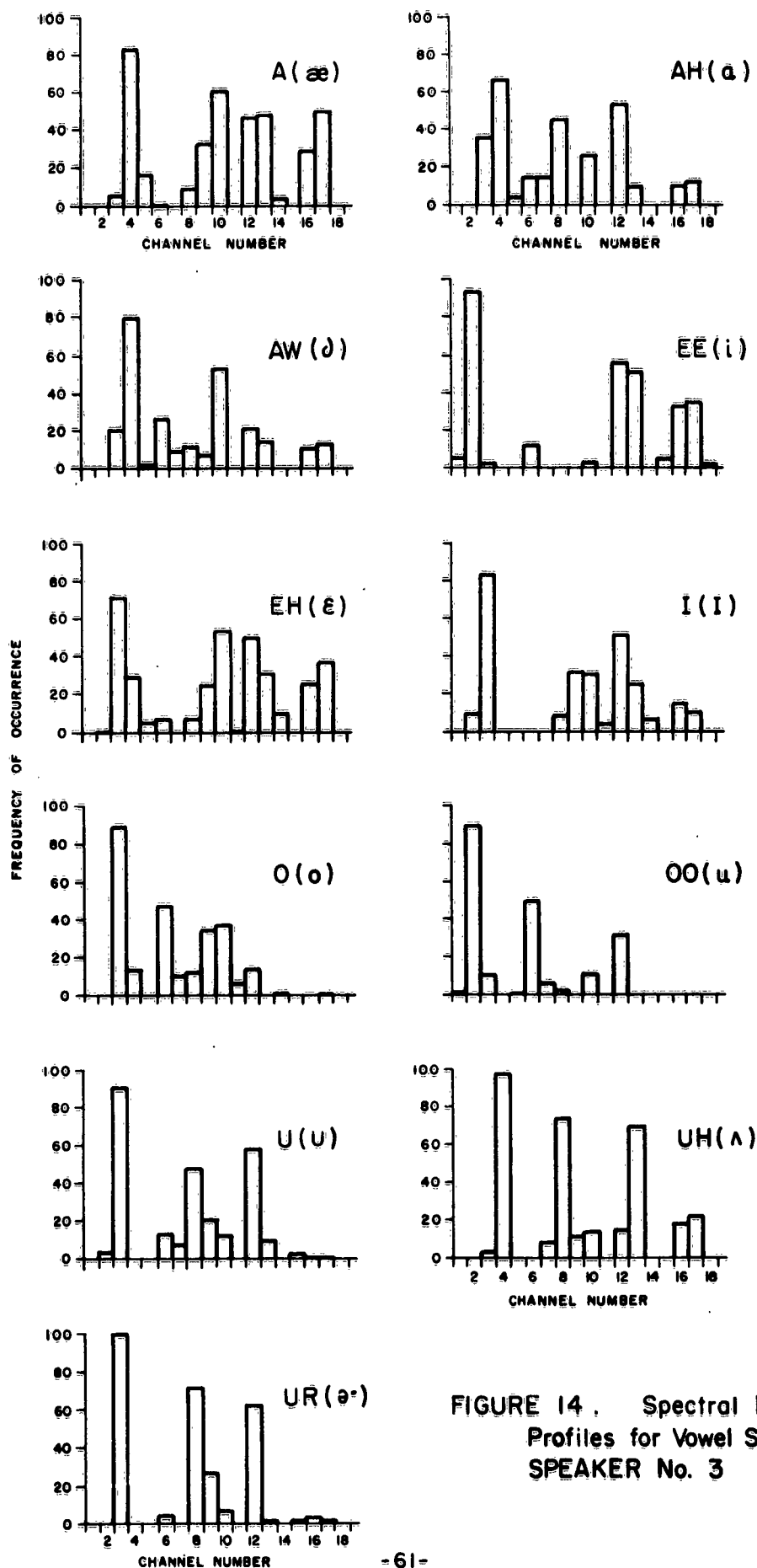


FIGURE 14. Spectral Peak Profiles for Vowel Sounds, SPEAKER No. 3

4. SPEECH TRANSCRIPTION AND WORD RECOGNITION TECHNIQUES

In this section, techniques for completing the final transformation, i. e. from parameter space to the space consisting of language elements, are discussed. The basic language element used to represent speech in this study is a speech "sound", as described in Section 2.1. In addition to methods of transforming patterns of parameter values into these speech sounds, techniques for recognizing spoken words as sequences of speech sounds are discussed. Illustrative examples of transcription quality obtainable with the simplest of these methods using a minimal parameter space (spectral peaks alone) are included at the end of this section.

4.1 SPEECH TRANSCRIPTION METHODS

It has been suggested in Section 2.3 that of all the different ways that sample patterns of parameter values could be associated with language elements, the maximum likelihood method (using histograms as estimates of probability distributions) seems to offer the greatest potential for good performance, if the number of speech samples used in constructing the histograms is large enough. Assuming that enough samples can be obtained (as is possible with the equipment described in Section 3.1), it might be concluded that only one step remains to complete the process of automatic transcription. This consists of implementing the table look-up operation dictated by the decision boundaries resulting from this maximum likelihood method of partitioning parameter space into nonoverlapping regions corresponding to different speech sounds. If essentially no overlap occurs in parameter space between speech sounds, this conclusion is correct. To produce a sequence of sound symbols representing speech, it is only necessary that each sample pattern of parameter values be compared with a collection of labeled patterns (called a "reference library"), and type the label corresponding to the reference pattern which is matched by the sample. Since most speech sounds (as defined on this project) span several speech samples, the occasional occurrence of no match between a single incoming sample pattern and any of the reference patterns will produce no significant loss of information. One way of handling no-match decisions is to produce a standard symbol, say "Y", indicating this fact; another option is to print out nothing for the no-match decisions. An idealized transcription of the word "THREE", using the latter option with the rudimentary exact match method applied to each speech sample, would be:

TH TH UR UR UR UR EE EE EE EE EE EE EE EE EE

In producing this transcription consisting of 15 sound symbols, perhaps 25 speech samples might have been processed, with 10 no-match decisions within the word. Although it may be desired that some indication be retained of the time intervals spanned by speech sounds (conceivably to identify the speaker by recreating an accent), it is anticipated that the most compact presentation would be desired for most applications. This can be achieved by modifying the rudimentary exact match method by printing out a sound symbol only if it is different from the preceding symbol. This modification produces "TH UR EE" for the above example. As with no-match decisions, any one of several methods can be employed to indicate samples taken during silence, or no-speech sounds, either indicating or not indicating the duration of such intervals.

If overlap exists between speech sounds in parameter space (i. e., if it is likely that a sizeable percentage of speech samples will be misclassified), then the rudimentary exact match method will produce a distorted transcription. Two avenues exist by which such a situation could be improved: (1) additional parameters can be extracted from speech signals, and (2) the way in which decisions are reached can be changed. The first approach is straightforward. Augmentation of peak-space with spectral moments, for instance, produces less overlap between vowel sounds, as indicated in the tables at the end of Section 3. As soon as enough parameters are available to produce separated speech sounds in parameter space, then the exact match transcription method may be employed as described above.

If it happens that not enough parameters can be used to achieve separation of speech sounds (without exceeding storage limitations, for instance), then the method of associating patterns of parameter values with speech sounds can be changed. The reason why there exists room for improvement over the single sample exact match method based on maximum likelihood is simply that sometimes several speech samples are taken within the intervals of speech corresponding to utterances of single speech sounds. It is therefore not necessary to render a decision for each speech sample. If some method is devised for segmenting speech into intervals corresponding to utterances of speech sounds, then all of the speech samples taken within each interval could be combined to produce a more reliable decision.

In any pursuit of this course for improving speech transcription quality, several methods of combining speech samples deserve investigation. Perhaps the most straightforward method consists of observing the sequence of sounds occurring in an interval (as determined by the rudimentary exact

match method of labeling each speech sample), and associating the speech segment with that sound which has occurred most frequently in the segment. This "plurality rule" method would require a relatively simple augmentation of the equipment required for the rudimentary exact match method alone. Another technique consists of selecting one of the samples occurring within the speech segment as a representative for that segment, and choosing the sound to which the representative is associated, with the maximum likelihood method applied to single samples. Two ways to select such a representative are (a) the segment midpoint, and (b) the sample occurring nearest to the point in the segment at which the speech signal is judged (by some operation) to be changing the least. The latter (quiescent) sample could be selected in a variety of ways.

When only one decision is to be rendered for each speech segment, it is also possible to combine the sequence of speech samples spanned by the segment to produce a single, derived parameter value which would represent the combination. This kind of operation would produce what might be called a derived parameter space consisting of a small number of dimensions. As an illustration of this approach, consider the sequence of s spectral peak patterns, $\underline{p}_i = (p_{i1}, p_{i2}, \dots, p_{i18})$, $i = 1, 2, \dots, s$, corresponding to a given speech segment. These samples may be combined to produce a single pattern, $\underline{u} = (u_1, u_2, \dots, u_{18})$, according to the formula:

$$u_j = \frac{1}{s} \sum_{i=1}^s p_{ij}, \quad j = 1, 2, \dots, 18.$$

The quantity u_j reflects the percentage of speech samples (with the given segment) which have a peak in the j -th frequency channel. If a sound produces mostly repetitions of the same peak pattern within a speech segment, then \underline{u} will be essentially identical with this peak pattern. If, on the other hand, a speech sound is characterized by slight changes in peak patterns within a speech segment, then \underline{u} will consist of some components which are less than one, but greater than zero. The amount and nature of the change in peak positions within the speech segment will be reflected by the shape of the pattern, \underline{u} .

For any given method of combining speech samples with a speech segment, there exist many methods of associating the resulting pattern \underline{u} , with speech sounds. As with individual speech samples, the method of maximum likelihood using histograms as estimates of the distribution of speech sounds in the derived parameter space, would probably provide the most accurate association. However, it is quite possible that the large

number of different patterns of derived parameter values which can result from utterances of speech sounds will preclude the collection of enough samples of speech to warrant the use of this method. It is likely, therefore, that one of the other methods described in Section 2.3 would have to be relied upon. With the method of combining peak patterns described above, for instance, the spectral profiles (Figures 12, 13, and 14) can be regarded as representatives of the speech sounds, and correlation between \underline{u} and a given spectral profile would provide an indication of "closeness" between the speech segment and the sound corresponding to the profile. The segment would be associated with the speech sound whose corresponding spectral profile produces the highest correlation with \underline{u} .

In order to implement any of these methods for combining several speech samples, a method of segmenting speech must be devised. From the little study of the parameter ΔS (Table 3) which could be conducted after its extraction was automatized toward the end of this project, it appears that speech signals can be partitioned into time intervals roughly corresponding to speech sounds by thresholding this quantity. As illustrated in Table 4, transitions between speech sounds can also be identified.

Before the efficacy of this or any other speech segmentation method can be ascertained, many experiments must be carried out using several of the more promising methods of combining speech samples. Time did not allow for such experimentation during this project. However, the exact match transcription method was programmed for simulation on the Recomp II computer, and several tests have been conducted. As described above, the rudimentary exact match method produces either a single phonetic symbol, or no symbol if a sample does not match any of the patterns stored in the reference library. Although tests were conducted for both peak space and peak-moment space, the number of no-match decisions obtained for the latter space precluded extensive study. The reasons for the large number of no-match decisions with the spectral moment parameters are twofold. First, instead of normalizing M_1 and M_2 with respect to M_0 , all three quantities were extracted and quantized separately. The quantization of M_0 thus produced considerable unnecessary variation in the measure of spectral spread, σ . Perhaps even more detrimental to proper extraction of moments, M_2 was quantized linearly, rather than logarithmically, thus producing high resolution for unvoiced spectra, but very coarse resolution for voiced spectra. Both of these problems were foreseen (and are easily remedied through the addition of modified analogue-to-digital converters in the experimental speech processing equipment), but could not be avoided within the time span of this project. The

transcriptions have therefore been conducted primarily for peak space only. For this parameter space, the random arrangement of errors in transcriptions using this method (with reference libraries constructed from data described in Section 3*), produced relatively long sequences of symbols. Although the correct speech sounds were represented more frequently than incorrect sounds, considerably study is required to make the identification. However, observation of the general persistence of the correct sounds over several samples suggested another option for "smoothing" the sequence of sounds, and combining several samples to produce a reduced number of symbols. Instead of typing out a single symbol for the most likely speech sound, if the two or three most likely sounds associated with each speech sample are selected as tentative candidates, and ambiguities are resolved in favor of sounds which persist as candidates over the largest number of samples, then fairly readable transcriptions are obtained. Specifically, the following procedure for processing rudimentary exact match transcriptions (with up to three candidates for each speech sample) has been followed:

- (1) Print out a sound symbol only if the same sound is recognized on two successive samples.
- (2) Repeat a sound symbol for every successive adjacent pair of occurrences of the sound.
- (3) Ambiguities are resolved in favor of the sound which either (a) has occurred on the previous sample, or (b) persists the longest without interruption.
- (4) Symbols for unvoiced sounds are inserted properly.

Examples of the resulting transcriptions obtained with this exact match and smoothing method** are shown in Tables 16, 17, and 18, for three different speakers. The rudimentary transcriptions were performed on the Recomp computer, and the smoothing operations were completed by hand.

*For peak space, as indicated in Table 9, the libraries consisted of 308 patterns for Speaker Number One, 343 patterns for Speaker Number Two, and 185 patterns for Speaker Number Three.

**From utterances of a test word list (Table 19).

TABLE 16. EXACT MATCH TRANSCRIPTION OF TEST WORD LIST FOR
SPEAKER NUMBER ONE (PEAK-SPACE)

<u>STANDARD TRANSCRIPTION</u>	<u>AUTOMATIC TRANSCRIPTION</u>
Z EE UR O	<u>Z</u> <u>II</u> UR O AH AH O
OO O AH N	OO OO U U/O U/O <u>N</u>
T U OO OO OO	<u>T</u> <u>II</u> OO OO
TH UR UR EE EE EE	<u>TH</u> OO UR/O EE EE EE EE EE EE
F O O O UR	<u>F</u> O O O O O AW UR/O
F AH A EH V	<u>F</u> UH UH UH UH UH UH UH OO <u>V</u>
S I I K S	<u>S</u> <u>I</u> <u>I</u> <u>K</u> <u>S</u>
S EH V U N	<u>S</u> I/UH/O <u>V</u> U/O <u>N</u>
EH EE T	<u>II</u> EE <u>T</u>
N AH EH I N	<u>N</u> O/U UR EE EE <u>N</u>
PL UH UH S	<u>P</u> <u>L</u> O AH UH <u>S</u>
M AH EH N U S	<u>M</u> U <u>II</u> EH <u>N</u> U <u>S</u>
T AH EH I M S	<u>T</u> EH EH U <u>I</u> <u>I</u> <u>M</u> <u>S</u>
P UR I N T	<u>P</u> UR I/EH <u>N</u> <u>T</u>
EE K OO U L S	EE EE EE <u>K</u> OO O O O O <u>L</u> <u>S</u>
ST AH UH P	<u>ST</u> UH UH UH <u>P</u>
P AW I N T	<u>P</u> O AW EE <u>N</u> <u>T</u>
ST AH UR T	<u>ST</u> UH UR <u>T</u>
A AW L F U	UH UH AW <u>L</u> <u>F</u> U UR
B EH EE T U	<u>B</u> EH/I EH/I <u>T</u> EH/U/UH
EH EH KS	AH/AW/UH EH <u>K</u> <u>S</u>
OO AH A EH I	OO/UR UH UH UR UH UH UH AW/EH/I
Z EE EE EE	<u>Z</u> EE EE EE EE EE EE
UR I P EE EE T	UR UR UR <u>P</u> EE EE EE <u>T</u>
TH UR OO OO OO	<u>TH</u> OO U OO OO OO

Notes: (1) Interpolated Sounds are Underlined.
(2) AH/EH indicates "either AH or EH".

TABLE 17. EXACT MATCH TRANSCRIPTION OF TEST WORD LIST FOR
SPEAKER NUMBER TWO (PEAK-SPACE)

<u>STANDARD TRANSCRIPTION</u>	<u>AUTOMATIC TRANSCRIPTION</u>
Z EE UR O	<u>Z</u> EE U U U U U
OO O AH N	UR UR AH <u>N</u>
T U OO OO OO	<u>T</u> U OO OO OO OO OO OO
TH UR UR EE EE EE	<u>TH</u> U U EE EE EE E
F O O O UR	<u>F</u> O O AH O O
F AH A EH V	<u>F</u> UH AH AW EH EH <u>V</u>
S I I K S	<u>S</u> I I I I <u>K S</u>
S EH V U N	<u>S</u> EH EH <u>V</u> U <u>N</u>
EH EE T	I I EE <u>T</u>
N AH EH I N	<u>N</u> UH/EH/A O EH EH <u>N</u>
PL UH UH S	<u>P</u> <u>L</u> AH AH UH <u>S</u>
M AH EH N U S	<u>M</u> AH/UH <u>N</u> EH I <u>S</u>
T AH EH I M S	<u>T</u> A EH EH <u>M S</u>
P UR I N T	<u>P</u> I <u>I N T</u>
EE K OO U L S	EE EE <u>K</u> U U U U <u>L S</u>
ST AH UH P	<u>S</u> <u>T</u> AH AH AH <u>P</u>
P AW I N T	<u>P</u> AW O I <u>N T</u>
S T AH UR T	<u>S</u> <u>T</u> AW/O AW/O UR/EH/I <u>T</u>
A AW L F U	AH AH <u>L</u> <u>F</u> O O
B EH EE T U	<u>B</u> I/UR EE <u>T</u> I I
EH EH K S	EH EH <u>K S</u>
OO AH A EH I	UR UR AH AH A A EH
Z EE EE EE	<u>Z</u> EE EE EE EE EE EE
UR I P EE EE T	EE <u>P</u> EE EE EE <u>T</u>
TH UR OO OO OO	<u>TH</u> OO U U U OO

Notes: (1) Interpolated Sounds are Underlined.
(2) AH/EH indicates "either AH or EH".

TABLE 18. EXACT MATCH TRANSCRIPTION OF TEST WORD LIST FOR
SPEAKER NUMBER THREE (PEAK-SPACE)

<u>STANDARD TRANSCRIPTION</u>	<u>AUTOMATIC TRANSCRIPTION</u>
Z EE UR O	<u>Z</u> EH UR UR UR UR O
OO O AH N	OO EH AW AW A/EH/I <u>N</u>
T U OO OO OO	<u>T</u> EE OO
TH UR UR EE EE EE	<u>TH</u> U U U U EE EE EE EE EE
F O O O UR	<u>F</u> O/AW O/AW O/AW O/AW AW AW
F AH A EH V	<u>F</u> AW AW AW A A A A EH EH <u>V</u>
S I I K S	<u>S</u> I I I <u>K</u> <u>S</u>
S EH V UN	<u>S</u> A A <u>V</u> EH EH EH <u>N</u>
EH EE T	EH/ I/UR EH/I/UR EE EE <u>T</u>
N AH EH IN	<u>N</u> EH AW AW AW AW EH/L/UR <u>N</u>
PL UH UH S	<u>P</u> <u>L</u> AW AW AW AW <u>S</u>
M AH EH N U S	<u>M</u> A A A EH EH <u>N</u> EH <u>S</u>
T AH EH I M S	<u>T</u> AW AW AW AW EH EH EH <u>M</u> <u>S</u>
P UR IN T	<u>P</u> UR I I <u>N</u> <u>T</u>
EE K OO U L S	EE EE EE <u>K</u> O O O O <u>L</u> <u>S</u>
S T AH UH P	<u>S</u> <u>T</u> AH/AW AH/AW AH/AW <u>P</u>
P AW IN T	<u>P</u> AW/EH AW/EH UR UR <u>N</u> <u>T</u>
S T AH UR T	<u>S</u> <u>T</u> EH EH A A A A <u>T</u>
A AW L F U	A/EH A/EH AH/AW A/AW <u>L</u> <u>F</u> AW O/U
B EH EE T U	<u>B</u> EH/I EH/I/UR EE T EH A A
EH EH KS	EH EH EH/I/UR <u>K</u> <u>S</u>
OO AH A EH I	OO AH A A AW AW A AW I I
Z EE EE EE	<u>Z</u> EE EE EE EE
UR I P EE EE T	EE EE <u>P</u> EE EE EE <u>T</u>
TH UR OO OO OO	<u>TH</u> U U OO OO

Notes: (1) Interpolated Sounds are Underlined.
(2) AH/EH indicates "either AH or EH".

The entire procedure can be instrumented quite easily.

Also shown in Tables 16, 17, and 18 is a "standard" transcription of the test word list obtained by a human transcriber after listening to several utterances of the words. It is clear that many different transcriptions would be equally acceptable and certainly possible as a result of variations in accent, as well as variations in interpretation by observers.

The sometimes perfect transcriptions obtained with this simple exact match transcription method, and using only peak patterns as the extracted parameters, suggests strongly that the addition of parameters reflecting spectral shape would produce highly readable transcriptions of all vowel sounds, and most voiced sounds.

4.2 WORD RECOGNITION METHODS

Although automatic transcription of speech into sequences of phonetic elements does not necessarily involve words as language elements at all, the possibility of using a speech transcriber for voice control of machines suggests that word recognition tests may afford a reasonable method of evaluating speech transcription methods. Although word recognition tests inherently involve not only the transcription methods, but the word recognition methods as well, we have adopted this method--as was suggested by the procuring agency.

To maximize the probability of correctly recognizing spoken words, it is probably true that decisions on the presence or absence of words should be based on intervals of observed speech which span the longest word in the given vocabulary. Furthermore, to maximize the information obtainable from an interval of observed speech for the purpose of deciding which (if any) of a given list of words has been spoken, no intermediate decisions should be made. From both of these standpoints, the recognition of phonetic elements as a preliminary to word recognition tends to degrade slightly the potential for achieving accurate word recognition for a given vocabulary. However, as pointed out previously, any attempt to utilize words as the basic language elements for transforming speech into readable text creates intolerable restrictions on the allowable speech which can be transformed, involves basic difficulties in changing vocabulary, and requires that initial decisions be rendered between a far larger number of alternatives--thus increasing equipment complexity significantly. Therefore, we must be content with achieving whatever performance is attainable through the use of sequences of sounds as the starting point for word recognition.

Two approaches have been considered for processing sequences of sounds to recognize words. With the first approach, a number is assigned to each sound in such a way that sequences of sounds corresponding to different words should be maximally differentiable from each other by the decision rule with which words are to be recognized. If, for instance, the word recognition method consists of correlating sequences of sounds with stored sequences, each of which represents a sound, then such a numerical assignment of numbers to sounds can have a relatively simple solution. Specifically, if all of the words to be recognized are so different as to produce uncorrelated sequences of sounds (if transcribed perfectly), then numbers should be assigned to sounds so that the variance of numbers corresponding to first sounds of all words in the vocabulary is maximized. Similarly, the variance of numbers associated with subsequent sounds in a perfect transcription should also be maximized.

The second approach to the assignment of numerical values to sounds is based on engineering considerations aimed at making the electronic implementation of word recognition particularly simple. Assume that sounds occurring in a specific word are assigned numerical values in agreement with the chronological sequence in which these sounds occur in the word. For instance, in the word "art", transcribed "AH UR T", if we assign numbers to the 3 different sounds so that AH = 1, UR = 2, and T = 3, then a rudimentary exact match transcription of the word "art" might appear as

AH	AH	AH	AH	UR	UR	UR	T	T
1	1	1	1	2	2	2	3	3

When associated with other words, the sounds AH, UR, and T may be assigned different numerical values so that, in the particular word in question, numbers assigned to sounds form a monotonically increasing sequence. This assignment is readily implemented by assigning to each sound recognition output unit (flip-flop), a voltage divider, where each tap on the divider is routed to different word-recognition units. Numerical values of voltages appearing at the taps correspond to the position of that sound in the sequence of sounds in the word to which the output of the tap is routed. Hence, as shown in Figure 15, the machine implemented by the above description consists of a number of different parts.

The machine will be described by reference to a specific example, wherein the recognition of only 2 words, the word "art" and the word "tar" are required. Ideal transcriptions of these two words contain three basic sounds. With the recognition of each there is associated a flip-flop labeled

FF(AH), FF(UR), FF(T). As a result of each speech sample, usually only one of the sound recognition flip-flops will be ON. It thus generates a pair of voltages at the two taps of the voltage dividers labeled AH_1 and AH_2 , UR_1 and UR_2 , T_1 and T_2 , depending on which flip-flop drives the attenuator. The subscripts indicate to which word recognition device (one or two, corresponding to the words "art" and "tar") the divider outputs are routed. The numerical values of the coefficients signify the position of the corresponding sound in the word whose number is denoted by the subscripts of the coefficient. Coefficients with like subscripts are added, resulting in the occurrence of a monotonic sequence of voltages at the output of that adder which corresponds to the word presently uttered. Since sound sequences corresponding to different words will not be identical for good transcriptions, only one of the summing devices will have a monotonic voltage output as a function of time. This will occur in the particular device which corresponds to the word being spoken.

The differentiator that follows each summing device will have an output which consists of a sequence of positive impulses if the right sequence of sounds, corresponding to the word of present interest, is uttered. Multiple successive occurrences of identical sounds will result in a differentiator output that still only consists of positive impulses, except that impulses will be missing at times corresponding to the multiple occurrence of identical sounds. The occurrence of negative impulses in any of the differentiator outputs indicates that the word corresponding to the particular summer differentiator probably did not occur. Recognition of a word should be based on a comparison of the numerical values of the output of a set of low-pass filters that follow the differentiators. The output of each low-pass filter is proportional to the number of positive impulses minus the number of negative impulses that occurred at the output of the differentiator over a period of time equal to the word length. Thus the word which resulted in the least number of errors in the expected sound sequence is said to have been spoken.

Since a thorough evaluation of either of these approaches to word recognition can be conducted only after a parameter space which separates essentially all speech sounds has been constructed, tests have been confined during this project to the easily simulated, engineering approach. A test word list consisting of 25 words has been selected, and exact match transcriptions have been used as inputs to 25 word recognition units designed in accordance with the illustration in Figure 15.

Several considerations have entered into the selection of a test vocabulary. First, for a specified vocabulary size, word recognition tends to be easier to perform if the length and variation in length of words are large. Therefore, to insure that a high level of difficulty is established for testing word recognition and transcription schemes, short test words should be selected for the test vocabulary.

For a given vocabulary size, word recognition tends to be less difficult if many different sounds are involved in the words. On the other hand, sound recognition may become more difficult as the number of allowable speech sounds is increased in a vocabulary. Since the test vocabulary is to be used both as a means of evaluating speech transcriptions and word recognition methods, we have chosen not to limit the sounds in a transcription to those involved in a test word vocabulary. Therefore, sound recognition capability is made independent of the test word vocabulary, and will not be affected by the distribution of sounds in these words. At the same time, we have chosen to select a vocabulary such that each word is not only short but tends to sound like a few of the other words in the vocabulary, so that word recognition, even by a human, may be a significantly difficult problem. The Test Word List (TWL) appears in Table 19.

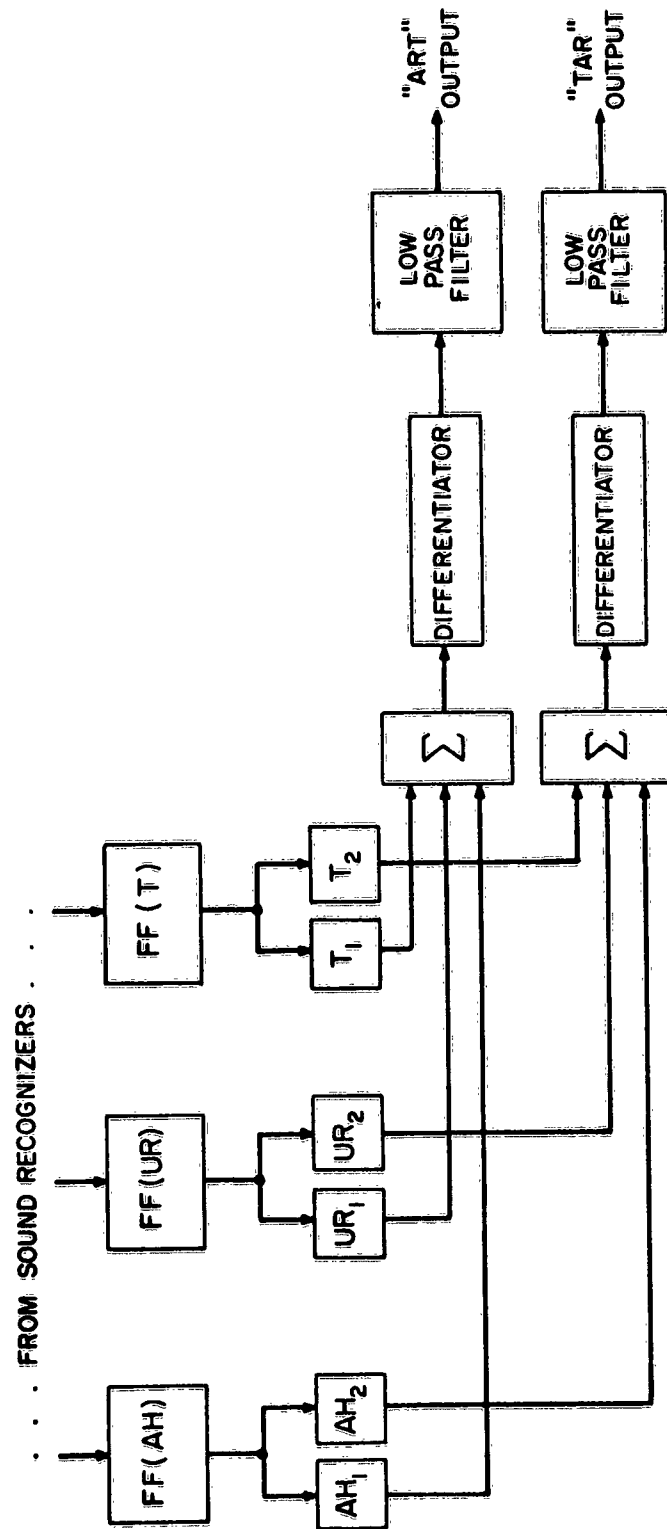
TABLE 19. TEST WORD LIST

ZERO	NINE	START
ONE	PLUS	ALPHA
TWO	MINUS	BETA
THREE	TIMES	X
FOUR	PRINT	Y
FIVE	EQUALS	Z
SIX	STOP	REPEAT
SEVEN	POINT	THROUGH
EIGHT		

This vocabulary was selected to illustrate typical commands and data for use in a computer-input application, as well as to satisfy the qualitative desiderata discussed above.

The ten spoken numerals have been used in the past* as test words. The other 15 words in the Test Word List include several word groups with common voiced and unvoiced sounds. The relative frequencies of occurrence

*For instance, [7], [11].



WP 62-1192

FIGURE 15. Recognition of the Words "ART" and "TAR"

of sounds in the TWL are shown in Table 20, along with the relative frequency of occurrence of sounds in conversational speech.* Although no major effort was made to match the distribution of sounds in the TWL precisely with their distribution in conversational speech, a close correspondence was obtained for all but six of the sounds. The 23 sounds involved in the TWL account for approximately 87 percent of sound occurrences in conversational speech.

To obtain a statistically significant indication of word recognition performance, five different utterances of each word in the test word list were transcribed with the rudimentary exact match method (interpolating non-vowel sounds), and each of the resulting 125 sequences of sounds were processed through each of 25 word recognition units. The transcriptions were performed by computer simulation, and the word recognition units were simulated by hand calculations. Approximately 80 percent correct identification of words was obtained, using only peak-patterns as the extracted parameters. While no tests were possible using spectral moments, as well as spectral peaks, it is anticipated that these simple exact match transcription and word recognition methods will produce greater than 90 percent correct identification of words, taking into account unvoiced, as well as voiced sounds.

*As derived from Table 15 in [4], p. 96.

TABLE 20. RELATIVE FREQUENCY OF OCCURRENCE OF SOUNDS

Sound		Relative Frequency of Sound Occurrence (Percent)	
Recomp	IPA	In Test Word List	In Conversational Speech*
AW	ɔ	1	1.3
OO	u	3	2.0
U	U	4	1.0
UR	ɔ'	4	3.0
AH	a,	8	3.5
UH	ʌ, 9	4	6.0
O	O, OU	2	1.7
A	a, ae	3	2.3
EH	ε	5	3.7
I	i	7	6.3
EE	i, j	6	6.3
L	l	3	4.6
R	r	3	3.1
W	w	2	3.7
M	m	2	3.6
N	n	7	8.1
NG	ŋ	0	1.1
B	b	1	0.6
D	d	0	4.6
G	g	0	1.5
Z	z	2	2.2
V	v	2	1.8
TJ	ʝ	0	2.5
ZH	ʒ	0	0.0
T	t	9	9.8
P	p	5	1.7
K	k	3	3.6
H	h	0	1.1
WH	hw	0	1.1
F	f	3	2.0
TH	θ	2	0.7
S	s	9	4.0
SH	ʃ	0	0.7
CH	tʃ	0	0.3
DJ	d	0	0.3

*From [4].

5. CONCLUSIONS AND RECOMMENDATIONS

The basic approach to speech transcription investigated on this project consists of (1) representing speech signals as sequences of periodic sample patterns of parameter values, called "instantaneous spectra", and (2) associating phonetic language elements with selected sets of patterns. To ascertain storage requirements and obtain estimates of the accuracy with which speech sounds can be represented, laboratory speech processing equipment (Figure 5) has been utilized to obtain data on several parameters (Table 9), and the representation of speech sounds in the parameter spaces constructed from two combinations of these parameters has been investigated. Methods of associating patterns of parameter values with speech sounds, and sequences of speech sounds with words, have also been examined. Although these methods were selected primarily on the basis of ease of instrumentation, they exhibit high potential for providing accurate transcriptions and word recognition. Salient conclusions and recommendations for further development of these methods are presented in the following paragraphs.

Accuracy

With respect to accuracy, Tables 12, 13, and 14 indicate that parameter spaces constructed from spectral peaks and a few other parameters reflecting spectral shape of speech samples can be expected to provide good separation of vowels and other voiced sounds. Specifically, the average estimated probability of correctly identifying the vowel sound from which a single 17 msec speech sample is taken, is approximately 0.74, using spectral peaks alone (peak space). Augmentation of peak space with the first two spectrum moments increases the estimated probability of correctly identifying a single vowel sample to 0.86. If the "plurality-rule" method (Section 4.1) of combining speech samples within segments corresponding to single speech sounds, is used to reduce the number of decisions rendered per unit time, then these individual sample probabilities could be expected to produce a probability of correct decision (for vowels) of 0.90 and 0.98, for peak-space and peak-moment space, respectively.*

* These figures are based on the assumption that an average of five speech samples occur within a speech segment.

Recognition of 25 or more words by processing sequences of transcribed speech sounds can be performed with relatively simple equipment (Figure 15). The accuracy attainable is expected to be quite high when several additional parameters are measured in conjunction with spectral peaks. Using spectral peaks alone with the rudimentary transcription method for vowel sounds and interpolating non-vowel sounds, an 80 percent probability of correct recognition of one of 25 words has been obtained with the most easily instrumented word recognition method.

Storage Requirements

The number of different patterns of parameter values which can occur in speech within an interval corresponding to a single decision serves as an indication of the efficiency with which speech signals are being processed, as well as the complexity of equipment required to render the decision automatically. With the rudimentary exact match method of associating a speech sound with each speech sample, the number of different patterns of parameter values is quite small. Using only spectral peaks in an 18 channel vocoder, for instance, there are less than 7000 different patterns which are possible. This would indicate that less than 13 bits of information are utilized for each decision. Moreover, taking into account the fact that not all possible patterns of parameter values are produced by speech signals, the information processed for each decision is even less. With spectral peaks, for instance, it is estimated (Section 3. 1) that no more than approximately 400 different spectral peak patterns would ever occur in vowel sounds; i. e. only 9 bits per decision would be required for vowel sounds. With the addition of other speech parameters the information storage requirements would increase, but evidently not drastically. With the addition of the first two spectral moments (properly quantized as indicated in Figure 9), it appears that three additional bits would suffice.

Implementation

From the standpoint of implementing an exact match transcription method, a reference library consisting of 1000 patterns can be handled quite easily. The exploitation of either "always" or "never" conditions for most of the binary quantities involved in patterns of parameter values, produces a decision "tree" with only a few nodes and branches. This transcription method can be implemented readily with diode matrices or relays.

Word recognition units can be constructed readily by the method indicated in Figure 15. It should be stressed that by first transcribing speech into sequences of speech sounds, essentially all restrictions on the number and type of different words which can be handled are lifted. Of course, performance will tend to be degraded as the number of words it is desired to distinguish between increases, but the construction of word recognition units can proceed independently of the transcription method being employed.

Recommendations

The data collection and analysis program reported here primarily for voiced sounds should be carried out for the remaining speech sounds. This would produce a complete indication of the transcription accuracy attainable with spectral peaks and spectrum moments.

Two courses for improving transcription accuracy, augmentation of parameters and modification of recognition methods (discussed in Section 4.1), should be pursued in the following way. First, additional speech parameters should be introduced to produce a parameter space in which all speech sounds are widely separated. In addition to normalization of the spectral moments, the following parameters deserve examination:

- (1) Derivative of Normalized Speech Envelope $\frac{dE_o}{dt}$
- (2) Silence Indication
- (3) Low Frequency First and Second Moments
- (4) High Frequency First and Second Moments
- (5) Duration of Unvoiced Intervals
- (6) Formant Time Derivative Polarity

With the addition of some of these parameters, the rudimentary exact match transcription method, with smoothing (see Section 4.1), should produce acceptable transcriptions for the majority of speech sounds.

To attain a readable transcription for all members of a phonetic alphabet, it may be necessary to introduce another method of recognition. Through the use of the parameter, ΔS (Section 3.1.1), speech may be segmented into short intervals corresponding to either utterances of speech sounds, or portions of speech sounds. By combining all of the patterns of parameter values occurring in a given segment, a more reliable decision can be rendered. As suggested in Section 4.1, several

methods of combining the samples occurring within a segment should be investigated thoroughly, including correlation of cumulative spectral peak counts with replicas of spectral profiles, and plurality rule of sounds within each segment.

APPENDIX I

Program for Simulating a Peak-Picking Formant Tracking Vocoder

The input to the program consists of a sequence of "instantaneous spectra" (samples of a vocoder output taken every Δ seconds) representing an isolated spoken word. The number of spectra in each utterance depends upon the duration of the word. Each spectrum is in 18 channel vocoder format with the energy in each channel quantized in 3 bits, and is represented by the quantities, a_1, \dots, a_{18} . The object of the program is to locate for each spectrum the frequency channels in which the energy exhibits a local maximum. The output for each spectrum consists of 18 bits, one for each vocoder channel, where a "one" indicates a peak and a "zero", no peak, in the corresponding channel. In addition, one bit for the voiced-unvoiced decision and three bits for the number of peaks are included. A flow for this program is shown in Figure 16.

The method of locating the local peaks may be described briefly as follows:

There is a peak in channel n if $a_n > a_{n+1}$ and $a_n > a_{n-1}$. a_0 and a_{19} are assumed equal to 0, to allow peaks at the ends. If there are several channels of equal magnitude surrounded by channels of smaller magnitude, there are two alternatives. If the number of equal channels is odd the peak is placed in the middle channel. If the number is even the middle lies between two channels. In this case the peak is placed on the side of the middle which has the largest surrounding channel; or if the two surrounding channels are equal the peak is placed arbitrarily on the low frequency side.

A result of the peak-picking operation is shown in Figure 17. The voiced-unvoiced decision is made using a linear discriminant.

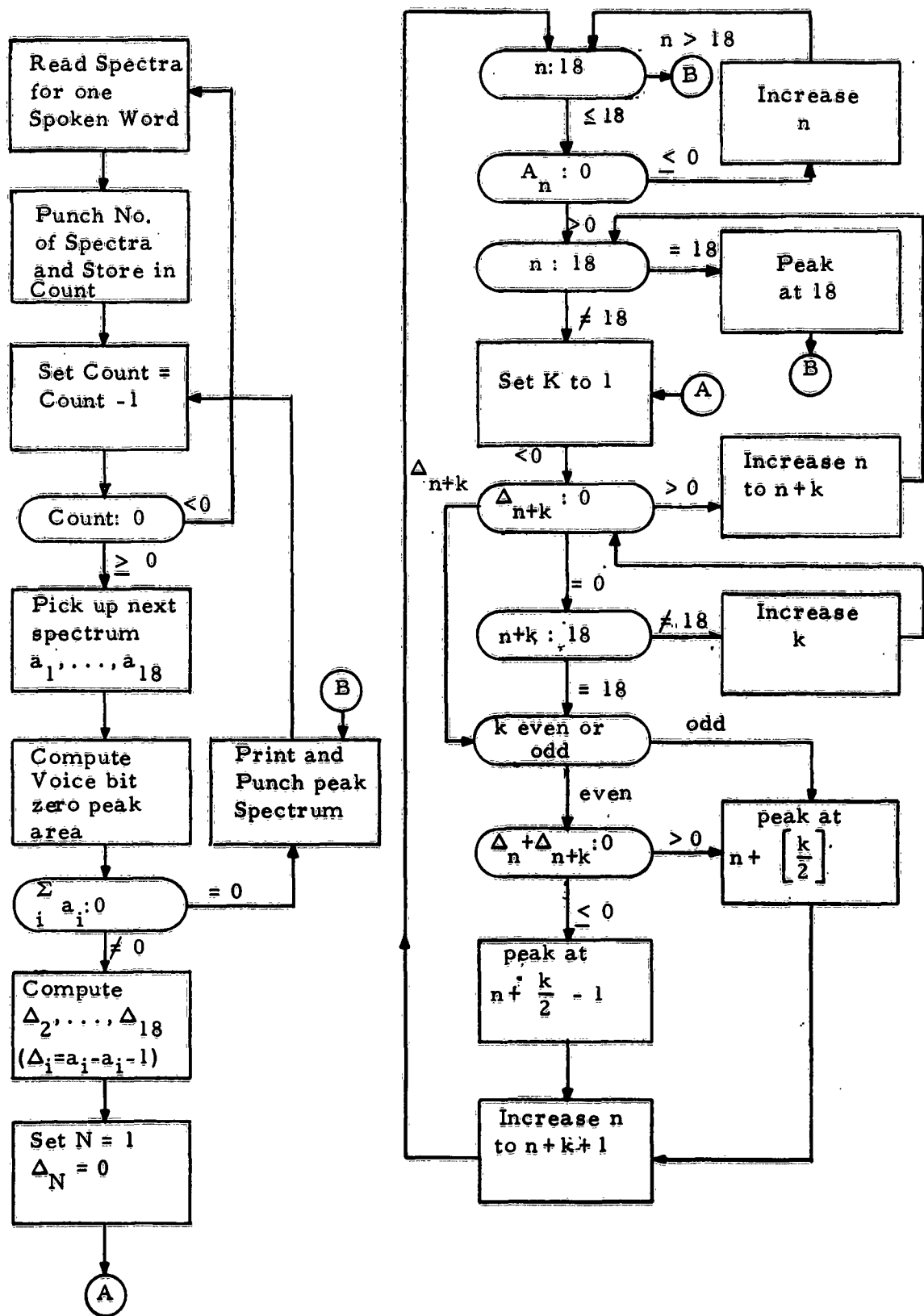


Figure 16. Flow Chart for Spectrum Peak Picking Program

<u>Original Spectrum</u>	<u>Peak Spectrum</u>	<u>Voiced</u>	<u>No. of Peaks</u>
11.1.....12311.	1..1.....1...	0	3
34.1.....1311.	.1.1.....1...	1	3
45111.....114211	.1.....1...	1	2
24222.....12621.	.1.....1...	1	2
113232.....12521.	..1.1.....1...	1	3
.12433.....11521.	...1.....1...	1	2
1113223....11252..	...1..1.....1...	1	3
.1133142...112421.	...1..1.....1...	1	3
...221441..111421.1.1.....1...	1	3
...22.242..222421.	...1..1.....1...	1	3
1..22.2341.22241..	1..1....1.....1...	1	4
1.13111141121331..	1..1....1..1.1....	1	5
21132.11411221211.	1..1....1..1..1...	1	5
52.11..13..221121.	1..1....1..1..1..	1	5
53.21...3..111121.	1..1....1.....1..	1	4
53.,1...31.121.21.	1...1...1..1..1..	1	5
53.....21.311.22.	1.....1..1..1..	1	4
53.....111.311.11.	1.....1..1..1..	1	4
63.....11.211.21.	1.....1..1..1..	1	4
52.....2..1.1.11.	1.....1..1.1.1..	1	5
62.....1.....	1.....1.....	1	2
6.....1.....	1.....1.....	1	2

Figure 17. Three-Bit Quantized and Peak-Picked Representation of the Spoken Word "ONE"

APPENDIX II

Ten Most Frequently Occurring Patterns in Peak Space and Peak-Moment Space, For each Vowel Sound and a Single Speaker

A. PEAK SPACE

The ten most frequently occurring patterns of values of the local spectral peaks are listed below for each of the eleven vowel sounds listed in Table 1.

Sound	Spectral Peaks																		Relative Frequency of Occurrence (%)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
A					1					1			1			1			21.3
					1					1			1						14.2
				1						1			1			1			12.3
					1					1			1			1		1	5.2
				1						1			1						4.8
					1					1			1					1	4.3
	1			1						1			1			1			4.3
					1					1		1				1			3.3
				1						1			1		1				2.8
				1					1			1							2.4

Sound	Spectral Peaks																		Relative Frequency of Occurrence (%)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
AH					1								1			1			14.9
					1									1		1			9.4
					1								1						8.8
					1									1					4.4
					1		1						1			1			4.4
	1				1					1			1			1			3.9
		1			1								1						3.2
			1		1					1						1			2.8
					1		1						1						2.2
	1			1									1			1			2.2

Sound	Spectral Peaks																		Relative Frequency of Occurrence (%)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
AW				1						1						1			20.0
				1									1			1			9.3
				1						1			1			1			5.1
					1					1			1			1			5.1
					1														4.7
				1												1			4.7
			1			1							1			1			4.7
				1		1				1			1						4.7
					1											1			4.2
						1								1		1			4.2

Sound	Spectral Peaks																		Relative Frequency of Occurrence (%)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
EE		1										1				1			16.4
			1													1			9.4
		1										1			1				9.4
			1									1			1				8.7
		1											1			1			7.0
		1													1	1			6.4
		1														1			4.7
		1										1			1			1	4.1
			1									1		1		1		1	4.1
		1				1										1			3.5

Sound	Spectral Peaks																		Relative Frequency of Occurrence (%)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
EH				1						1			1			1			15.6
				1						1			1			1			7.8
				1						1				1		1			7.8
					1					1			1						4.9
					1					1			1			1			4.9
				1						1			1						3.9
				1						1				1					3.9
					1							1				1			2.9
					1					1		1							1.9
					1					1			1						1.9

Sound	Spectral Peaks																		Relative Frequency of Occurrence (%)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
i			1							1			1			1			24.9
			1						1				1			1			7.7
			1							1			1						4.8
		1						1		1			1						3.8
			1					1		1			1						3.8
			1							1		1				1			3.8
				1						1			1						3.8
			1							1			1				1		2.9
				1						1		1				1			2.9
				1						1			1			1			2.9

Sound	Spectral Peaks																		Relative Frequency of Occurrence (%)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
O			1						1										11.0
			1																8.8
			1						1				1			1			6.6
			1					1											5.2
				1			1						1			1			4.4
							1						1						2.9
							1									1			2.9
							1						1			1			2.9
				1				1											2.2
			1						1				1						2.2

Sound	Spectral Peaks																		Relative Frequency of Occurrence (%)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
OO			1																30.6
			1					1											16.5
		1																	8.8
			1			1													8.8
		1				1													7.1
			1					1											4.7
			1													1			2.4
		1						1								1			2.4
		1	1			1													1.8
			1																1.2

Sound	Spectral Peaks																		Relative Frequency of Occurrence (%)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
U			1						1				1			1			12.8
			1						1			1				1			11.8
			1						1			1			1				5.9
			1					1											4.9
			1							1			1			1			4.9
			1																3.9
			1						1			1							2.9
			1						1							1			2.9
			1										1			1			2.9
				1				1									1		2.9

Sound	Spectral Peaks																		Relative Frequency of Occurrence (%)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
UH				1					1				1			1			11.5
				1						1			1			1			10.8
			1						1				1			1			8.3
			1				1						1			1			5.8
				1				1					1			1			5.8
					1			1				1							4.1
					1				1				1			1			3.3
					1		1			1			1			1			3.3
					1		1						1			1			2.5
					1			1					1			1			2.5

Sound	Spectral Peaks																		Relative Frequency of Occurrence (%)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
UR			1						1										26.8
			1							1									7.9
				1				1											7.3
			1					1		1									7.3
			1					1											6.1
			1						1							1			4.9
				1					1										4.3
			1						1				1						4.3
			1				1			1									3.7
		1							1										3.1

B. Peak-Moment Space

The ten most frequently occurring patterns of values of the local spectral peaks and the first three spectral moments are listed below for each of the eleven vowel sounds listed in Table 1.

Sound	Spectral Peaks																		M ₀	M ₁	M ₂
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18			
A					1					1			1			1			111	100	010
				1						1			1			1			110	011	001
					1					1			1			1			110	100	010
					1					1			1						110	011	010
				1						1			1			1			111	100	010
					1					1			1			1			111	101	011
					1					1			1						111	100	010
					1				1			1							110	011	010
					1					1			1						111	101	011
					1					1			1						110	011	001

Sound	Spectral Peaks																		M ₀	M ₁	M ₂
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18			
AH					1							1				1			101	011	001
					1							1							110	011	001
					1							1				1			110	011	010
					1							1		1		1			110	100	010
					1		1					1				1			111	100	010
	1				1					1		1				1			111	100	010
					1							1							101	011	001
					1							1				1			110	100	010
					1							1		1		1			111	100	010
				1								1				1			111	100	010

Spectral Peaks																					
Sound	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	M ₀	M ₁	M ₂
AW				1						1						1			1 1 0	0 1 1	0 1 0
				1									1			1			1 1 1	1 0 0	0 1 0
					1														1 1 0	0 1 1	0 0 1
				1						1			1						1 1 0	0 1 1	0 0 1
				1						1						1			1 1 1	0 1 1	0 1 0
				1						1						1			1 1 0	0 1 1	0 0 1
				1						1						1			1 1 1	1 0 0	0 1 0
				1								1							1 1 0	0 1 1	0 0 1
				1												1			1 0 1	0 1 0	0 0 1
				1												1		1 1 0	1 0 0	0 1 0	

Spectral Peaks																			M ₀	M ₁	M ₂
Sound	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18			
EE	1											1				1			1 0 0	0 1 0	0 1 0
		1														1			1 0 1	0 1 1	0 1 0
	1											1			1				1 0 1	0 1 1	0 1 0
		1										1			1				1 0 1	1 0 0	0 1 1
	1				1													1	1 0 1	1 0 0	0 1 1
	1															1			1 0 1	0 1 1	0 0 1
		1														1			1 0 0	0 1 1	0 0 1
	1															1			1 0 1	0 1 1	0 1 0
	1											1				1			1 1 0	0 1 1	0 1 0
1											1				1			1 1 0	1 0 0	0 1 0	

Spectral Peaks																					
Sound	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	M ₀	M ₁	M ₂
EH				1						1			1			1			1 1 1	1 0 0	0 1 0
			1							1			1			1			1 1 1	1 0 0	0 1 0
			1							1				1		1			1 0 1	0 1 0	0 0 1
				1						1			1			1			1 1 1	1 0 1	0 1 1
					1					1			1			1			1 1 1	1 0 0	0 1 1
					1						1		1			1			1 1 1	1 1 0	1 0 0
				1						1			1			1			1 1 1	1 0 0	0 1 1
				1									1			1			1 1 1	1 0 1	0 1 1
			1							1			1			1			1 1 1	1 0 1	0 1 1
		1								1				1		1			1 1 1	1 0 0	0 1 0

Spectral Peaks																					
Sound	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	M ₀	M ₁	M ₂
U			1						1			1			1				110	011	001
			1						1				1				1		110	011	010
			1						1			1							110	011	001
			1						1								1		110	010	001
			1						1			1					1		111	100	010
			1						1				1				1		111	011	010
			1						1				1				1		111	011	010
								1											110	010	001
														1			1		011	000	000
			1					1									1		110	011	001
			1						1				1						110	011	001

Spectral Peaks																					
Sound	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	M ₀	M ₁	M ₂
UH			1					1					1				1		111	100	010
				1				1					1				1		111	100	010
				1					1				1				1		111	100	010
				1						1			1				1		111	100	010
			1				1						1				1		111	101	011
			1					1					1				1		110	011	010
				1					1				1				1		110	011	001
				1						1			1				1		111	100	011
				1							1		1				1		111	011	010
				1				1					1					111	101	011	

Spectral Peaks																					
Sound	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	M ₀	M ₁	M ₂
UR			1						1										1 1 0	0 1 0	0 0 1
			1						1										1 1 0	0 1 1	0 0 1
				1				1											1 1 0	0 1 1	0 0 1
			1					1		1									1 1 0	0 1 0	0 0 1
			1					1											1 1 0	0 1 1	0 0 1
			1				1			1									1 1 0	0 1 1	0 0 1
			1					1											1 1 0	0 1 0	0 0 1
				1					1										1 1 0	0 1 1	0 0 1
			1						1				1						1 1 1	0 1 1	0 0 1
			1						1								1		1 1 1	0 1 1	0 1 0

APPENDIX III

A Program for Mapping Peak-Picked Spectra into a Reduced Space

The purpose of this program is the simultaneous generation of an "intermediate reference library" of "instantaneous spectra" in the 18 bit peak-picked format, and recording of speech data as a sequence of intermediate reference library numbers.

This program, designated as SMREF, sets up a library of reference patterns for speech data based on the following rules for similarity of two input spectra:

1. Voiced-unvoiced designation must be the same.
2. The number of peaks must be the same.
3. If the number of peaks is zero or one, the spectra must be identical.
4. If the number of peaks is greater than one but less than seven, corresponding peaks of one spectrum must not be more than one channel away from those of the other and the direction of the shift in peak locations must be the same.

Input is a series of tapes; the first record in each section indicates the number of vectors to follow, where each vector is a one word record describing the peak patterns, i. e., the location of the peaks, the voicing indication and the number of peaks. The input is compared against all previously established reference patterns. If a match is found, the "matching count" for the reference is up-dated. If no match, the input is stored as a new reference pattern. For each input the number of the matching reference spectrum is typed.

After all input spectra have been examined, the library is sorted and a three sectional tape is punched. The first contains the unvoiced patterns arranged according to number of peaks followed by the "count" of unvoiced patterns. The second consists of the same data for the voiced sounds. Section three is the unsorted reference library and all necessary controls for continuing the library generation at a later date. A copy of the sorted libraries is also typed.

There are the following restrictions:

1. Program is designed for eighteen channel data with a maximum of six peaks.
2. There are approximately 3,000 (decimal) locations reserved for the reference pattern library. If an extraordinary number of input vectors is used, there is a possibility of exceeding this space. (Loc. 0045.1 indicates the storage location for the storage location for the next reference pattern. This should not exceed 6777.0).

The program has been written for the Recomp II Computer for Contract AF30(602)-2641, February, 1962.

Flow charts for this program are shown in Figures 18, 19, and 20.

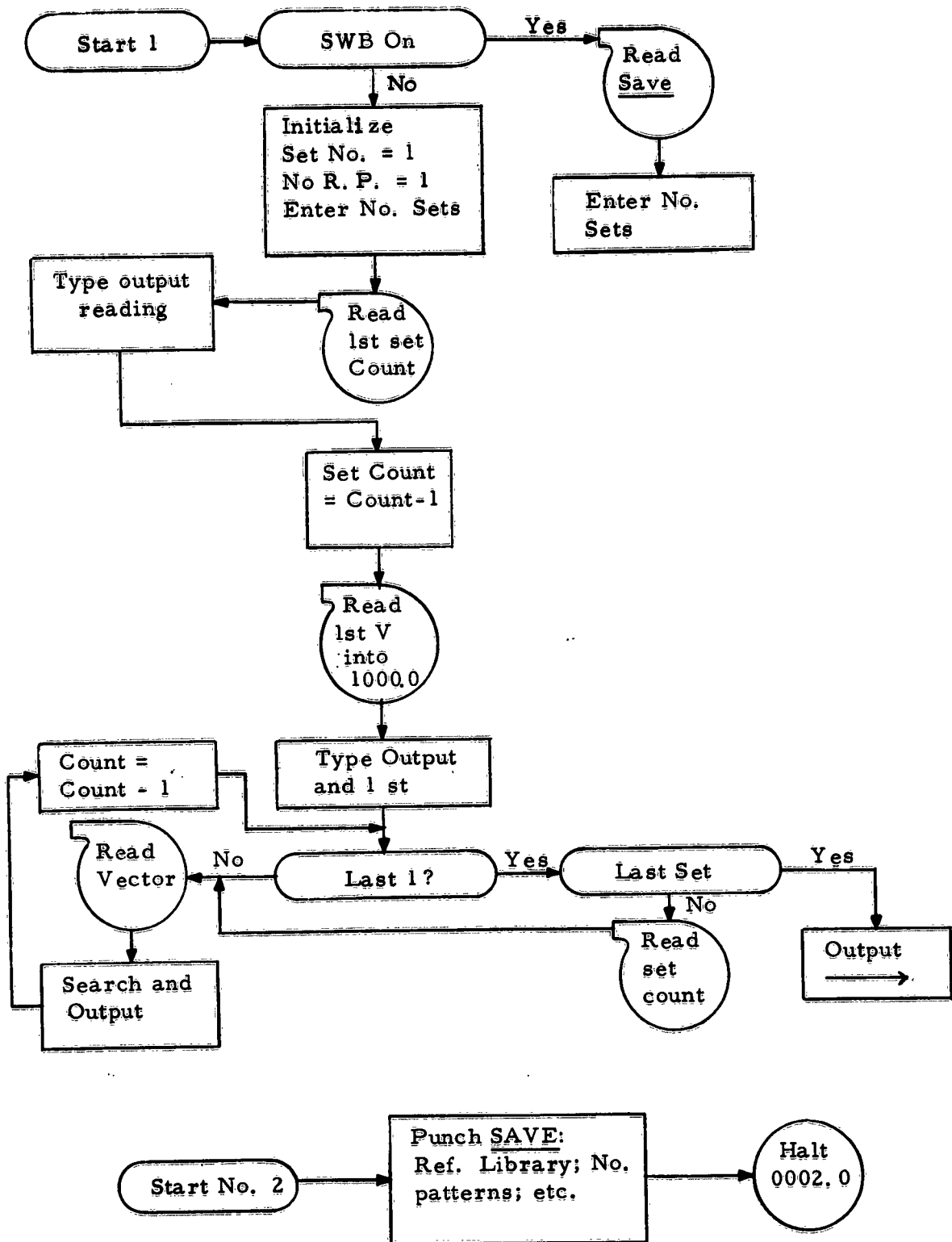


Figure 18. Flow Chart for SMREF

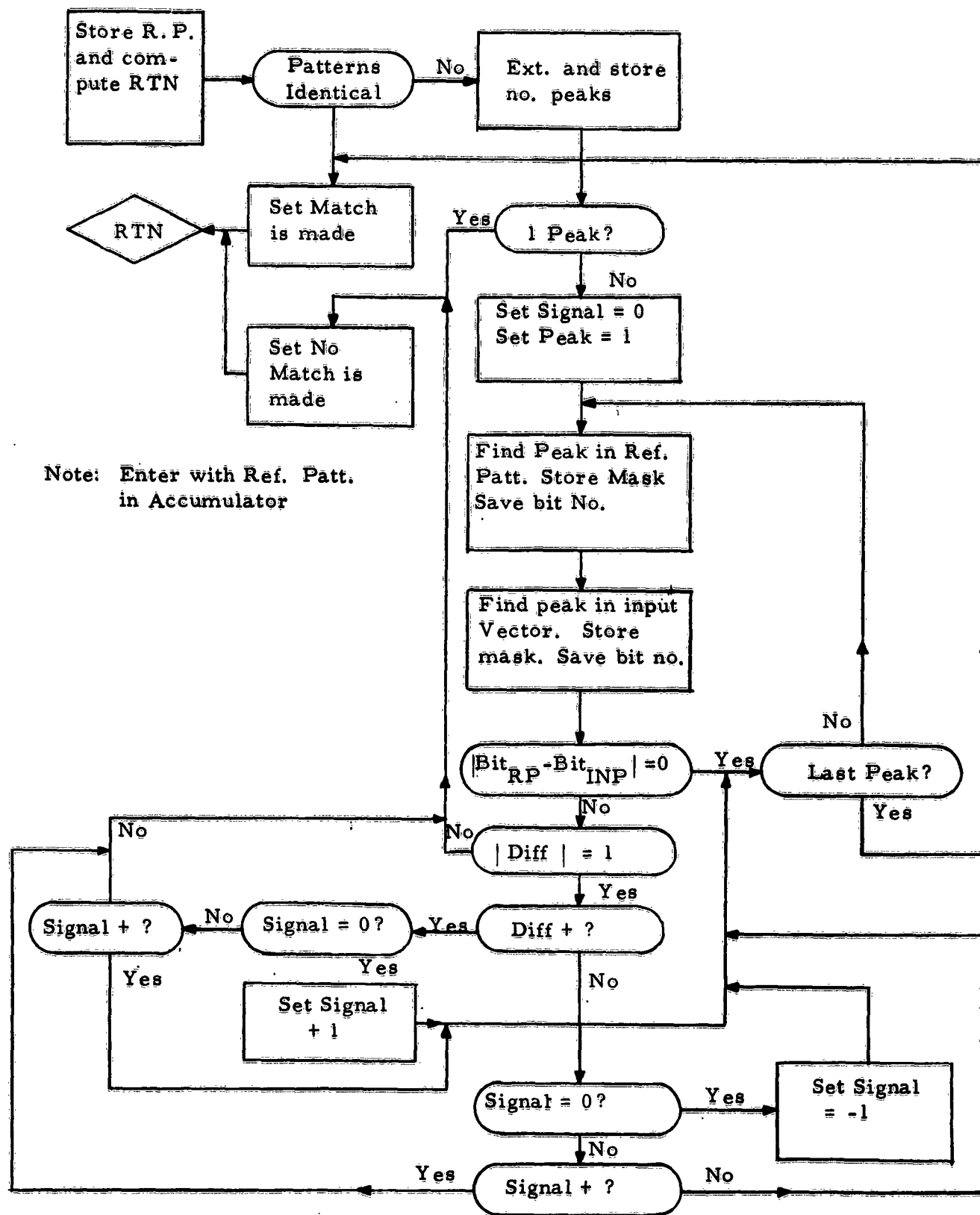


Figure 19. Flow Chart for Matching (Subroutine for SMERF)

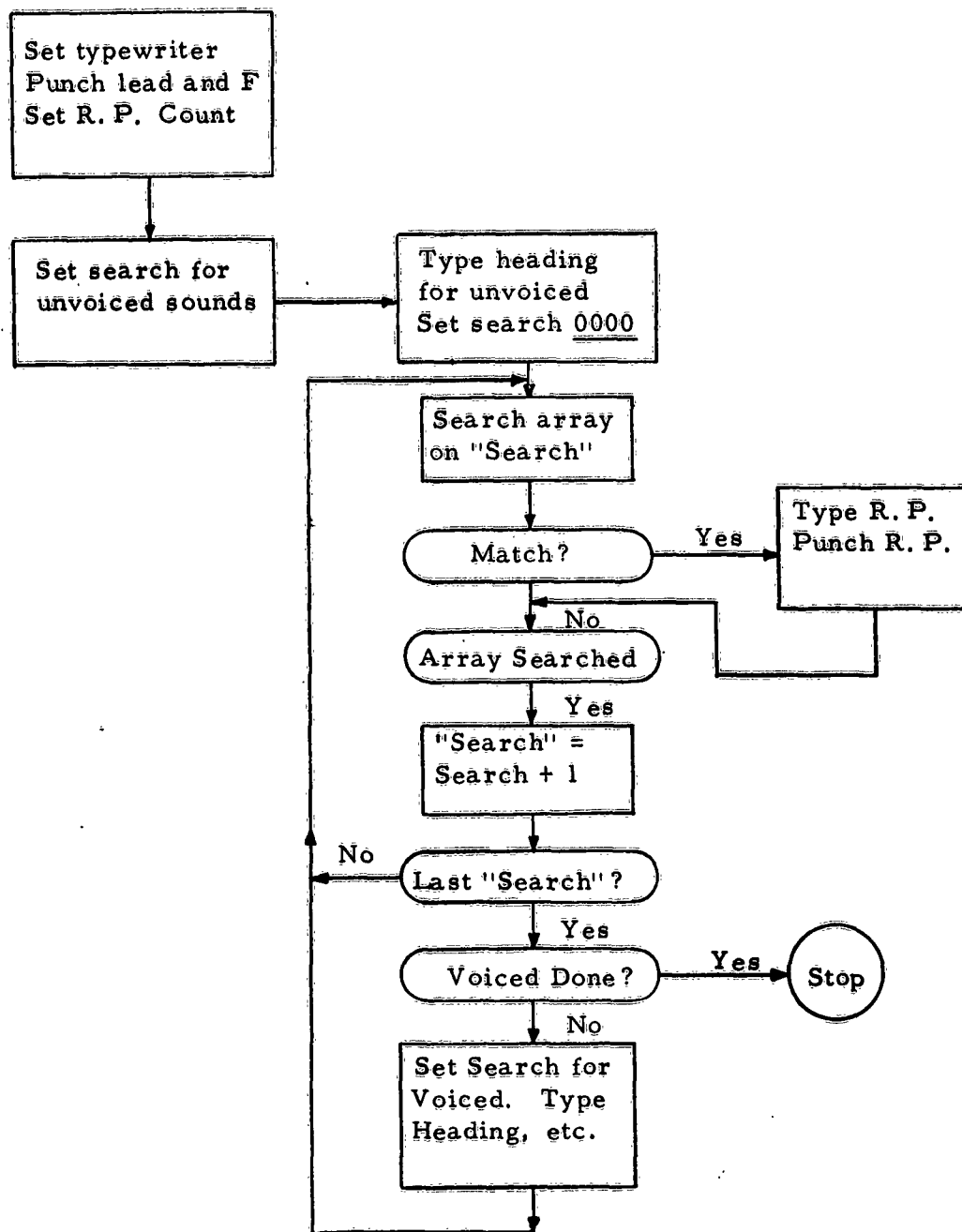


Figure 20. Output Section for SMREF

List of References

1. Cherry, E. C., "Roman Jakobson's Distinctive Features as the Normal Coordinates of Language", Mouton and Company, The Hague, 1956.
2. Fant, C. G. M., Speech Analysis and Synthesis, Summary Report Contract AF61(052)-342, January 31, 1961.
3. Fletcher, H., Speech and Hearing in Communication, D. Van Nostrand Co., Inc., 1965.
4. Gleason, H. A., An Introduction to Descriptive Linguistics, Holt, Rinehart and Winston, New York, 1961.
5. Hughes, G. W., "The Recognition of Speech by Machine", RLE Technical Report No. 395, MIT, May 1, 1961.
6. Olson, H. F., and Belar, H., "Phonetic Typewriter III", J. Acoust. Soc. Am., 33, 1610, (1961).
7. Sebestyen, G., "Recognition of Membership in Classes", Trans. IRE, PGIT, January, 1961.
8. Sebestyen, G., and Hartley, A., "Pattern Recognition Research", Final Report on Contract No. AF19(604)-8024, 31 December 1961.
9. Sebestyen, G., and Van Meter, D., "Investigation of Automation of Speech Processing for Voice Communication", Scientific Report on Contract No. AF19(604)-8828, May 28, 1962.
10. Sebestyen, G., Decision-Making Processes in Pattern Recognition, Macmillan, 1962.
11. Sholtz, P. N., and Bakis, R., "Spoken Digit Recognition Using Vowel Consonant Segmentation", J. Acoust. Soc. Am., 34, 1, (1962).
12. "US News and World Report", LIII, No. 8, p. 10, August 20, 1962.